Multivariate classification with random forests for gravitational wave searches of black hole binary coalescence

Paul T. Baker,[1] Sarah Caudill,[2,*] Kari A. Hodge,[3] Dipongkar Talukder,[4] Collin Capano,[5] and Neil J. Cornish[1]

[1]*Montana State University, Bozeman, Montana 59717, USA*
[2]*Leonard E. Parker Center for Gravitation, Cosmology, & Astrophysics, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201, USA*
[3]*California Institute of Technology, Pasadena, California 91125, USA*
[4]*University of Oregon, Eugene, Oregon 97403, USA*
[5]*Maryland Center for Fundamental Physics & Joint Space Science Institute, Department of Physics, University of Maryland, College Park, Maryland 20742, USA*

(Received 19 December 2014; published 13 March 2015)

Searches for gravitational waves produced by coalescing black hole binaries with total masses $\gtrsim 25\ M_\odot$ use matched filtering with templates of short duration. Non-Gaussian noise bursts in gravitational wave detector data can mimic short signals and limit the sensitivity of these searches. Previous searches have relied on empirically designed statistics incorporating signal-to-noise ratio and signal-based vetoes to separate gravitational wave candidates from noise candidates. We report on sensitivity improvements achieved using a multivariate candidate ranking statistic derived from a supervised machine learning algorithm. We apply the random forest of bagged decision trees technique to two separate searches in the high mass ($\gtrsim 25\ M_\odot$) parameter space. For a search which is sensitive to gravitational waves from the inspiral, merger, and ringdown of binary black holes with total mass between $25\ M_\odot$ and $100\ M_\odot$, we find sensitive volume improvements as high as $70_{\pm 13}\%$–$109_{\pm 11}\%$ when compared to the previously used ranking statistic. For a ringdown-only search which is sensitive to gravitational waves from the resultant perturbed intermediate mass black hole with mass roughly between $10\ M_\odot$ and $600\ M_\odot$, we find sensitive volume improvements as high as $61_{\pm 4}\%$–$241_{\pm 12}\%$ when compared to the previously used ranking statistic. We also report how sensitivity improvements can differ depending on mass regime, mass ratio, and available data quality information. Finally, we describe the techniques used to tune and train the random forest classifier that can be generalized to its use in other searches for gravitational waves.

DOI: 10.1103/PhysRevD.91.062004 PACS numbers: 04.80.Nn, 07.05.Kf, 07.05.Mh

I. INTRODUCTION

We are rapidly approaching the era of advanced gravitational-wave detectors. Advanced LIGO [1] and Advanced Virgo [2] are expected to begin operation in 2015. Within the next decade, these will be joined by the KAGRA [3] and LIGO-India [4] detectors. The coalescence of compact binaries containing neutron stars and/or stellar mass black holes are expected to be a strong and promising source for the first detection of gravitational waves [5]. Higher mass sources with total masses $\gtrsim 25\ M_\odot$ including binary black holes (BBHs) and intermediate mass black holes (IMBHs) are less certain but still potentially strong sources [5–7]. Discovery and new science will be possible with detection of gravitational waves from these objects [8,9].

Measurement of gravitational waves requires exquisitely sensitive detectors as well as advanced data analysis techniques [10]. By digging into detector noise for weak signals rather than waiting for a rare, loud event, we increase detection rates. Unfortunately, detector noise can be nonstationary and non-Gaussian, leading to loud, short duration noise transients. Such behavior is particularly troublesome for higher mass searches where the expected in-band signal is of similar duration as noise transients. Traditional searches for compact binary coalescence have utilized multidetector coincidence, carefully designed ranking statistics, and other data quality methods [6,7,11–14]. However, in many searches performed to date over initial LIGO and Virgo data, the sensitivity was limited by an accidental coincidence involving a non-Gaussian transient noise burst [6,7,14].

Only recently have gravitational-wave searches begun to utilize methods that work with the full multidimensional parameter space of classification statistics for candidate events. Previous studies have shown multivariate methods give detection probability improvement over techniques based on single parameter thresholds [15–17].

Machine learning has a wealth of tools available for the purpose of multivariate statistical classification [18,19]. These include but are not limited to artificial neural networks [20,21], support vector machines [22,23], and random forests of decision trees [24]. Such methods have already proven useful in a number of other fields with large

*Corresponding author.
sarah.caudill@ligo.org

data sets and background contamination including optical and radio astronomy [25–27] and high energy physics [28,29]. Within the field of gravitational wave physics, a search for gravitational-wave bursts associated with gamma ray bursts found a factor of ∼3 increase in sensitive volume when using a multivariate analysis with boosted decision trees [30]. Applications of artificial neural networks to a search for compact binary coalescence signals associated with gamma ray bursts found smaller improvements [31]. Machine learning algorithms have successfully been applied to the problem of detector noise artifact classification [32]. Additionally, a search for bursts of gravitational waves from cosmic string cusps [33] used the multivariate technique described in [15].

In this paper, we focus on the development and sensitivity improvements of a multivariate analysis applied to matched filter searches for gravitational waves produced by coalescing black hole binaries with total masses $\gtrsim 25\ M_\odot$. In particular, we focus on the application to two separate searches in this parameter space. The first, designated the inspiral, merger, and ringdown (IMR) search, looks for gravitational waves from the inspiral, merger, and ringdown of BBHs with total mass between $25\ M_\odot$ and $100\ M_\odot$. The second, designated the ringdown-only search, looks for gravitational waves from the resultant perturbed IMBH with mass roughly between $10\ M_\odot$ and $600\ M_\odot$. These investigations are performed over data collected by LIGO and Virgo between 2009 and 2010 so that comparisons can be made with previous IMR and ringdown-only search results [6,7]. Using a random forest of bagged decision trees (RFBDT) supervised machine learning algorithm (MLA), we explore sensitivity improvements over each search's previous classification statistic. Additionally, we describe techniques used to tune and train the RFBDT classifier that can be generalized to its use in other searches for gravitational waves.

In Sec. II, we frame the general detection problem in gravitational-wave data analysis and motivate the need for multivariate classification. In Sec. III, we describe our data set. In Sec. IV, we explain the method used to classify gravitational-wave candidates in matched-filter searches. In Sec. V, we review RFBDTs as used in these investigations. In Sec. VI, we discuss the training set, the multidimensional space used to characterize candidates, and the tunable parameters of the classifier. In Sec. VII, we describe the improvement in sensitive volume obtained by the IMR and ringdown-only searches over LIGO and Virgo data from 2009 to 2010 when using RFBDTs. Finally, in Sec. VIII we summarize our results.

II. THE DETECTION PROBLEM

Searches for gravitational waves are generally divided based on the astrophysical source. The gravitational waveform from compact binary coalescence has a well-defined model [34,35]. Thus searches for these types of signals use the method of matched filtering with a template bank of model waveforms. This is the optimal method for finding modeled signals with known parameters buried in Gaussian noise [36,37]. However, if the parameters are not known, matched filtering is not optimal [38], and additional techniques must be employed to address the extraction of weak and/or rare signals from non-Gaussian, nonstationary detector noise, the elimination or identification of false alarms, and the ranking of gravitational-wave candidates by significance. This paper presents the construction of an *ad hoc* statistic, automated through machine learning, that can tackle these issues.

A. Searches for compact binary coalescence

The coalescence of compact binaries generates a gravitational-wave signal composed of inspiral, merger, and ringdown phases [34,35]. The low frequency inspiral phase marks the period during which the compact objects orbit each other, radiating energy and angular momentum as gravitational waves [39]. The signal for low mass systems in the LIGO and Virgo frequency sensitivity bands (i.e., above the steeply rising seismic noise at 40 Hz for initial detectors or 10 Hz for advanced detectors [40]) is dominated by the inspiral phase. Several searches have looked for the inspiral from low mass systems including primordial black hole binaries with component masses in the range $0.2\ M_\odot$–$1.0\ M_\odot$ [41] and neutron star and/or black hole systems with component masses $>1\ M_\odot$ and total mass $<25\ M_\odot$ [11–13]. The higher frequency merger phase marks the coalescence of the compact objects and the peak gravitational-wave emission [42–44]. Since the merger frequency is inversely proportional to the mass of the binary, the signal for high mass systems in the LIGO and Virgo sensitivity bands could include inspiral, merger, and ringdown phases. Searches for high mass signals including all three phases have been performed for systems with total mass between $25\ M_\odot$ and $100\ M_\odot$ [6,14]. We designate this as the IMR search. Systems accessible to LIGO and Virgo with even higher total masses will only have a ringdown phase in band, during which the compact objects have already formed a single perturbed black hole [45,46]. Searches for ringdown signals have looked for perturbed black holes with total masses roughly in the range $10\ M_\odot$–$600\ M_\odot$ and dimensionless spins in the range 0 to 0.99 [7,47]. The dimensionless spin is defined as $\hat{a} = cS/GM^2$ for black hole mass M and spin angular momentum S. We designate this as the ringdown-only search.

Each of these searches uses a matched-filter algorithm with template banks of model waveforms to search data from multiple gravitational-wave detectors. The output is a signal-to-noise ratio (SNR) time series for each detector. We record local maxima, called triggers, in the SNR time series that fall above a predetermined threshold. Low mass searches use template banks of inspiral model waveforms

generated at 3.5 post-Newtonian order in the frequency domain [48,49]. These waveforms typically remain in the initial LIGO/Virgo frequency sensitivity band for tens of seconds providing a natural defense against triggers arising from short bursts of non-Gaussian noise.

The templates for IMR searches include the full inspiral-merger-ringdown waveform, computed analytically and tuned against numerical relativity results. For these investigations, the nonspinning EOBNRv1 family of IMR waveforms was used [50]. The templates, like those for the low mass search, are described by the chirp mass $\mathcal{M} = \eta^{3/5}M$ and symmetric mass ratio $\eta = m_1 m_2/M^2$ of the component objects (where $M = m_1 + m_2$) [51]. The duration of high mass waveforms in band for the initial detectors is much shorter than the duration for low mass waveforms, making the IMR search susceptible to triggers associated with short bursts of non-Gaussian noise.

The templates for the search for perturbed black holes, with even higher total mass, is based on black hole perturbation theory and numerical relativity. A perturbed Kerr black hole will emit gravitational waves in a superposition of quasinormal modes of oscillation characterized by a frequency $f_{\ell mn}$ and damping time $\tau_{\ell mn}$ [45,52]. Numerical simulations have demonstrated that the $(\ell, m, n) = (2, 2, 0)$ dominates the gravitational-wave emission [46,53]. From here on, we will designate f_{220} as f_0 and write the damping time τ_{220} in terms of the quality factor $Q = Q_{220} = \pi f_{220}\tau_{220}$. Ringdown model waveforms decay on the time scale $0.0001 \lesssim \tau/\mathrm{s} \lesssim 0.1$, again making this search susceptible to contamination from short noise bursts.

The matched-filter algorithms are described in [47,51]. Further details on the templates and template bank construction in the IMR and ringdown-only searches can be found in [6,7].

Matched filtering alone cannot completely distinguish triggers caused by gravitational waves from those caused by noise. Thus tools such as data quality vetoes, multidetector coincidence, and SNR consistency checks are needed [54–62]. Additionally, a χ^2 time-frequency signal consistency test augments searches with a broadband signal including the IMR search but is less useful for short, quasimonochromatic ringdown signals [63]. Finally, each search uses a detection statistic to summarize the separation of signal from background. Details on the construction of a detection statistic are provided in Sec. IV.

In general, coincidence tests are applied to single detector triggers to check for multidetector consistency. The low and high mass searches use an ellipsoidal coincidence test (*ethinca* [60]) that requires consistent values of template masses and time of arrival. The ringdown-only search coincidence test similarly calculates the distance ds^2 between two triggers by checking simultaneously for time and template coincidence ($\mathrm{d}f_0$ and $\mathrm{d}Q$) [7]. When three detectors are operating, if each pair of triggers passes the coincidence test, we store a triple coincidence. We also store double coincidences for particular network configurations as outlined in Sec. III.

B. Signal and background

Evaluating the performance of a detection statistic and training the machine learning classifier require the calculation of detection efficiency at an allowed level of background contamination. In the absence of actual gravitational-wave events, we determine detection efficiency through the use of simulated signals ("injections") added to the detectors' data streams. To estimate the search background, we generate a set of accidental coincidences using the method of time-shifted data.

The simulated signal set is added to the data and a separate search is run. Triggers are recorded corresponding to times when injections were made. The simulated signals are representative of the expected gravitational waveforms detectable by a search. For the IMR and ringdown-only searches, the simulated signals include waveforms from the EOBNRv2 family [64] for systems whose component objects are not spinning and from the IMRPhenomB family [65] for systems whose component objects have aligned, antialigned, or no spins. Additionally, for the ringdown-only search, we inject ringdown-only waveforms. For a discussion of injection waveform parameters, see Sec. VI B 1. Considerations for the injection sets used in training the classifier are discussed in Sec. VI B, and those used in computing search sensitivity are discussed in Sec. VII.

The background rate of accidental trigger coincidence between detectors is evaluated using the method of time-shifted data. We shift the data by intervals of time longer than the light travel time between detectors and then perform a separate search. Any multidetector coincidence found in the time-shifted search is very likely due to non-Gaussian glitches. We performed searches over 100 sets of time-shifted data and recorded the accidental coincidences found by the algorithm. Details of the method are provided in Sec. III B of [6] and Sec. III C of [7]. For a discussion of the set of accidental coincidences used in training the classifier, see Sec. VI B.

III. DATA SET

We performed investigations using data collected by the LIGO and Virgo detectors between July 2009 and October 2010 [66]. Following the naming convention set in [7], we designate this time as Period 2 to distinguish it from the analysis of Period 1 data collected between November 2005 and September 2007. All results reported here consider only Period 2 data.

Period 2 covers LIGO's sixth science run [67]. During this time, the 4 km detectors in Hanford, Washington (H1), and in Livingston, Louisiana (L1), were operating. The 3 km Virgo detector (V1) in Cascina, Italy, conducted its

second and third science runs during this time [68]. The investigations were performed using data from the coincident search networks of the H1L1V1, H1L1, H1V1, and L1V1 detectors. Coincidences were stored for all triple and double detector combinations.

Data were analyzed separately using the IMR and the ringdown-only search pipelines from the analyses reported in [6] and [7]. In order to combat noise transients, three levels of data quality vetoes are applied to remove noise from LIGO-Virgo data when searches are performed. Details of vetoes are provided in [54–59] and specific descriptions of the use of the vetoes for these Period 2 analyses can be found in [6,7,17].[1] We analyze the performance of RFBDT classification after the first and second veto levels have been applied and compare this to the performance after the first, second, and third veto levels have been applied. After removal of the first and second veto levels, we were left with 0.48 years of analysis time and after the additional removal of the third veto level, we were left with 0.42 years of analysis time. We designate the search over Period 2 after the removal of the first and second veto levels as Veto Level 2 and after the removal of the first, second, and third veto levels as Veto Level 3.

In order to capture the variability of detector noise and sensitivity, we divided Period 2 into separate analysis epochs of ∼1 to 2 months. We then estimated the volume to which each search was sensitive by injecting simulated waveforms into the data and testing our ability to recover them. The sensitive volume that we compute in Sec. VII is found by integrating the efficiency of the search over distance,

$$V = 4\pi \int \eta(r) r^2 dr, \tag{1}$$

where the efficiency $\eta(r)$ is calculated as the number of injections found with a lower combined false alarm rate (refer to Sec. IV) than the most significant coincident event in each analysis epoch divided by the total number of injections made at a given distance. Details of the method are provided in Sec. V of [6,7]. We repeat this procedure in Sec. VII in order to quantify the improvement in sensitive volume obtained by each search over LIGO and Virgo data from 2009 to 2010 when using RFBDTs.

IV. DETECTION STATISTICS

We must rank all coincidences based on their likelihood of being a signal. Gravitational-wave data analysis has no dearth of statistics to classify gravitational-wave candidates as signal or background, and often, the ranking statistic will be empirically designed as a composite of other statistics. If the noise in the detector data was Gaussian, a matched-filter SNR would be a sufficient ranking statistic. However, since detector noise is non-Gaussian and nonstationary, we often reweight the SNR by additional statistics that improve our ability to distinguish signal from background. The exact form will depend on the nature of the signal for which we are searching. A good statistic for differentiating long inspirals may not work well for short ringdowns.

Searches for low mass binaries have ranked candidates using matched-filter SNR weighted by the χ^2 signal consistency test value (e.g., *effective* SNR [11] and *new* SNR [13]). IMR searches have used similar statistics [6,69,70]. Previous ringdown-only searches and studies have used SNR-based statistics to address the non-Gaussianity of the data without the use of additional signal-based waveform consistency tests (e.g., the chopped-L statistic for double [47] and triple [71,72] coincident triggers). However, ranking statistics can be constructed using multivariate techniques to incorporate the full discriminatory power of the multidimensional parameter space of gravitational-wave candidates. Several searches have utilized this including [30,33]. In Sec. V, we detail the implementation of a multivariate statistical classifier using RFBDTs for the most recent IMR and ringdown-only searches.

The final statistic used to rank candidates in order of significance is known as a detection statistic. We combine the ranking statistics for each trigger into a coincident statistic (i.e., a statistic that incorporates information from all detectors' triggers found in coincidence). This coincident statistic is then used to calculate a combined false alarm rate, the final detection statistic of the search. We determine the coincident statistic R for three different types of coincidences: gravitational-wave search candidates, simulated waveform injection coincidences, and time-shifted coincidences. We then determine a false alarm rate (FAR) for each type of coincidence by counting the number of time-shifted coincidences found in the analysis time T for each of the coincident search networks given in Sec. III. For each of the different types of coincidences in each of the search networks, we determine the FAR with the expression

$$\mathrm{FAR} = \frac{\sum_{k=1}^{100} N_k(R \geq R^*)}{T}, \tag{2}$$

where N_k is the measured number of coincidences with $R \geq R^*$ in the kth shifted analysis for a total of 100 time-shifted analyses. We then rank coincidences by their FARs across all search networks into a combined ranking, known as combined FAR [73]. This is the final detection statistic used in these investigations.

V. MACHINE LEARNING ALGORITHM

In order to compute the FAR for any candidate, we use the RFBDT algorithm to assign a probability that any candidate is a gravitational-wave signal. This random forest technology is a well-developed improvement over the classical decision tree. Each event is characterized by a

[1]The naming convention for veto categorization can vary across searches. We use the convention for Veto Levels 1, 2, and 3 as defined in Sec. V of [55].

feature vector, containing parameters thought to be useful for distinguishing signal from background. A decision tree consists of a series of binary splits on these feature vector parameters. A single decision tree is a weak classifier, as it is susceptible to false minima and overtraining [74] and generally performs worse than neural networks [29].

Random forest technology combats these issues and is considered more stable in the presence of outliers and in very high dimensional parameter spaces than other machine learning algorithms [19,74]. We use the StatPatternRecognition (SPR) software package[2] developed for high energy physics data analysis. This analysis uses the Random Forest of Bootstrap AGGregatED (bagged) Decision Trees algorithm. The method of bootstrap aggregation, or "bagging" as described below, tends to perform better in high noise situations than other random forest methods such as boosting [75].

A. Random forests

A random forest of bagged decision trees uses a collection of many decision trees that are built from a set of training data. The training data are composed of the feature vectors of events that we know *a priori* to belong to either the set of signals or the background (i.e., coincidences associated with simulated injections and coincidences associated with time-shifted data). Then the decision trees are used to assign a probability that an event that we wish to classify belongs to the class of signal or background. A cartoon diagram is presented in Fig. 1.

To construct a decision tree, we make a series of one-dimensional splits on a random subset of parameters from the feature vector. Each split determines a branching point, or node. Individual splits are chosen to best separate signal and background given the available parameters. There are several methods to determine the optimal parameter threshold at each node. We measure the goodness of a threshold split based on receiver operating characteristic curves (as described in Sec. VI A). The SPR software package [74] provides several options for optimization criterion. We found that the Gini index [76] and the negative cross entropy provided comparable, suitable performance for both searches. Thus we arbitrarily chose the Gini index for the IMR search and the negative cross entropy for the ringdown-only search. Additional discussion is given in Sec. VI D 4. The Gini index is defined by

$$G(p) = -2p\bar{p}, \tag{3}$$

where p is the fraction of events in a node that are signals and $\bar{p} = 1 - p$ is the fraction of events in the node that are background. Splits are made only if they will minimize the Gini index. The negative cross-entropy function is defined by

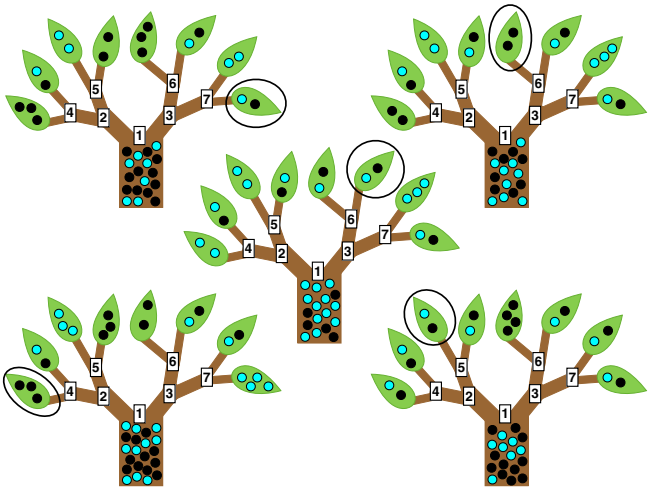


FIG. 1 (color online). A cartoon example of a random forest. There are five decision trees in this random forest. Each was trained on a training set of objects belonging to either the black set or the cyan set. Note that the training set of each decision tree is different from the others. At each numbered node, or split in the tree, a binary decision based on a threshold of a feature vector parameter value is imposed. The decisions imposed at each node will differ for the different trees. When no split on a node can reduce the entropy or it contains fewer events than a preset limit, it is no longer divided and becomes a leaf. Consider an object that we wish to classify as black or cyan. Suppose the object ends up in each circled leaf. Then the probability that the object is black is the fraction of black objects in all leaves, $p_{\text{forest}} = 73\%$.

$$H(p) = -p\log_2 p - \bar{p}\log_2 \bar{p}. \tag{4}$$

As H is symmetric against the exchange of p and $\bar{p}$, if a node contains as many signal events as background, then $H(\frac{1}{2}) = 1$. A perfectly sorted node has $H(1) = H(0) = 0$. By minimizing the negative cross entropy of our nodes, we find the optimal sort.

When no split on a node can reduce the entropy or it contains fewer events than a preset limit, it is no longer divided. At this point, the node becomes a "leaf." The number of events in the preset limit is known as the "minimum leaf size." When all nodes have become leaves, we have created a decision tree. This process is repeated to create each decision tree in the forest. Each individual tree is trained on a bootstrap replica, or a resampled set of the original training set, so that each tree will have a different set of training events. Furthermore, a different randomly chosen subset of feature vector parameters is chosen at each node to attempt the next split. Thus each decision tree in the forest is unique. Results from each tree can be averaged to reduce the variance in the statistical classification. This is the method of bootstrap aggregation or bagging.

The forest can then be used to classify an event of unknown class. The event is placed in the initial node of each tree and is passed along the various trees' branches until it arrives at a leaf on each tree. To compute the ranking

[2] http://statpatrec.sourceforge.net.

statistic for an event from a forest of decision trees, we find its final leaf in all trees. The probability that an event is a signal is given by the fraction of signal events in all leaves,

$$p_{\text{forest}} = \frac{\sum s_i}{\sum s_i + b_i} = \frac{1}{N}\sum s_i, \tag{5}$$

where s_i and b_i are the number of signal and background events in the ith leaf and N is the total number of signal and background events in all final leaves. The final ranking statistic, M_{forest}, for a forest of decision trees is given by the ratio of the probability that the event is a signal to the probability that the event is background,

$$M_{\text{forest}} = \frac{p_{\text{forest}}}{1 - p_{\text{forest}}}. \tag{6}$$

This ranking statistic is a probability ratio, indicating how much the signal model is favored over the background model.

In order to test the performance of the random forest, we determine how well it sorts events that we know *a priori* are signal or background. Rather than generate a new training set of simulated injections and time-shifted data, we may sort the training set used in construction of the forest. To prevent overestimation of classifier performance, decision trees cannot be used to classify the same events on which they were trained. Thus we use a round-robin approach to iteratively classify events using a random forest trained on a set excluding those events. We construct ten random forests each using 90% of the training events such that the remaining 10% of events may be safely classified. In this way we can efficiently verify the performance of the random forest using only the original training events.

Each forest is trained on and used to evaluate a particular type of double coincidence from the detector network (i.e., H1L1, H1V1, L1V1), as each pair of detectors produces unique statistics. Triple coincidences are split into their respective doubles, as there is not sufficient triple-coincident background to train a separate forest. For a triple coincidence with triggers from detectors a, b, and c, the ranking statistic $M_{\text{forest,triple}}$ will be the product of $M_{\text{forest,double}}$ for each pair of triggers,

$$M_{\text{forest,triple}} = \prod_i M_{\text{forest},i}, \tag{7}$$

where $i \in \{ab, ac, bc\}$ denotes the possible pairs of double coincident triggers in a triple coincidence.

VI. TUNING

There are several issues to consider when optimizing the performance of RFBDT classifiers. Performance of the algorithm is dependent on the quality of the training set (i.e., how well the training data represent the actual population we wish to detect). Additionally, we must select appropriate statistics to include in the feature vector of each coincidence. Finally, RFBDT classifiers have several parameters that must be tuned to optimize the sorting performance. These include number of trees, number of sampled parameters from the feature vector at each node, and minimum leaf size. Improperly choosing these meta parameters will lead to a poorly trained classifier. A summary of the tuned parameters used for each search is given in Table I.

In general there are two types of mistrained classifiers. An overtrained classifier separates the training data well, but the sort is specific to those data. An overtrained classifier may provide very high or very low M_{forest}, but these values contain a large systematic bias. Any events that were not well represented by this training set will be misclassified. An undertrained classifier has not gleaned enough information from the training data to sort properly. In this case the classifier is unsure of which set any event belongs, assigning intermediate values of M_{forest} to all.

A. Figure of merit

We evaluate the performance of different tunings using receiver operating characteristic (ROC) curves separately for each search. In general, these curves show the detection rate as a function of the contamination rate as a discriminant function is varied. For our purposes, thresholds on the combined FAR serve as the varying discriminant function. Thus, both the detection and the contamination rates are functions of the combined FAR.

Since we seek to improve the sensitivity of our searches, we reject the traditional definition of detection rate and instead define a quantity that depends on a distance-cubed weighting for each found injection. This quantity is essentially a fractional volume computed at each combined FAR threshold,

$$\frac{V_{\text{found}}}{V_{\text{all}}} = \frac{\sum_i \epsilon_i r_i^3}{\sum_i r_i^3}, \tag{8}$$

where i sums over all injections recovered as coincidences by the analysis pipeline, and r_i is the physical distance of the injection. Thus, V_{found} is the sum of volumes defined by r_i for each found injection and V_{all} is the sum of volumes

TABLE I. Summary of random forest parameters.

Search	IMR	Ringdown only
Number of trees	100	2000
Minimum leaf size	5	65
Total number of parameters	15	24
Number of randomly sampled parameters per node	6	14
Criterion for optimization	Gini index	cross entropy

TABLE II. Summary of full coalescence signal training set. These injections are parametrized by total binary mass M, mass ratio q, and mass-weighted spin parameter χ_s.

	EOBNRv2		IMRPhenomB	
Search	IMR	Ringdown only	IMR	Ringdown only
Mass distribution:	Uniform in (m_1, m_2)	Uniform in (M, q)	Uniform in (M, q)	Uniform in (M, q)
Total mass range:	$M/\mathrm{M}_\odot \in [25, 100]$	$M/\mathrm{M}_\odot \in [50, 450]$	$M/\mathrm{M}_\odot \in [25, 100]$	$M/\mathrm{M}_\odot \in [50, 450]$
Mass ratio range:	$q \in [1, 10]$	$q \in [1, 10]$	$q \in [1, 10]$	$q \in [1, 10]$
Spin parameter distribution:	Nonspinning	Nonspinning	Uniform in χ_s	Uniform in χ_s
Spin parameter range:	$\chi_s = 0$	$\chi_s = 0$	$\chi_s \in [-0.85, 0.85]$	$\chi_s \in [0, 0.85]$

defined by r_i for all injections. For each combined FAR threshold λ^*, ϵ_i counts whether injection i was found with a combined FAR λ_i less than or equal to λ^*,

$$\epsilon_i = \begin{cases} 1: & \lambda_i \leq \lambda^* \\ 0: & \lambda_i > \lambda^* \end{cases}.$$

In the following sections, we explore tunings and performance for the RFBDT algorithm for different total masses and mass ratios as well as at Veto Levels 2 and 3 in order to understand how the application of vetoes affects the RFBDTs.

B. Training set

The training of the classifier utilizes the signal and background data sets as described in Sec. II B. In the following discussions, we consider several issues that arise in the construction of training sets for gravitational-wave classification when using RFBDTs.

1. Signal training

In order to train the classifier on the appearance of the signal, we injected sets of simulated waveforms into the data and recorded those found in coincidence by the searches.

Both searches injected sets of waveforms from the EOBNRv2 and IMRPhenomB families. The total mass M, mass ratio q, and component spin $\hat{a}_{1,2}$ distributions of these waveforms are given in Table II. We define the mass ratio as $q = m_>/m_<$ where $m_> = \max(m_1, m_2)$ and $m_< = \min(m_1, m_2)$. The component spins are $\hat{a}_{1,2} = cS_{1,2}/Gm_{1,2}^2$ for the spin angular momenta $S_{1,2}$ and masses $m_{1,2}$ of the two binary components. From this, we define the mass-weighted spin parameter

$$\chi_s = \frac{m_1\hat{a}_1 + m_2\hat{a}_2}{m_1 + m_2}. \quad (9)$$

Additionally, for the ringdown-only search, we injected two sets of ringdown-only waveforms as described in Table III. The two sets gave coverage of the ringdown template bank in (f_0, Q) space and of the potential $(M_f, \hat{a})$ space accessible to the ringdown-only search where M_f is the final black hole mass and $\hat{a}$ is the dimensionless spin parameter. All injections were given isotropically distributed sky location and source orientation parameters. As described below, only injections that are cleanly found by the search algorithm are used in training the classifiers.

For performance investigations in Sec. VII, we determine search sensitivities using all injections found by the searches' matched filtering pipelines (i.e., not just those that are cleanly found). The IMR search considers full coalescence waveforms from the EOBNRv2 and IMRPhenomB families. The ringdown-only search considers only EOBNRv2 waveforms. These injection sets and their parameters are given in Sec. VII.

To identify triggers associated with simulated waveform injections made into the data, we use a small time window of width ± 1.0 s around the injection time. We record the parameters of the trigger with the highest SNR within this time window and associate it with the injection. Unfortunately, when injections are made into real data containing non-Gaussian noise, the injection may occur near a non-Gaussian feature or glitch in the data. In the case where the SNR of the injection trigger is smaller than that of the glitch trigger, the recorded trigger will correspond to the glitch trigger and will not accurately represent the simulated waveform. When using injections to train the classifier on the appearance of gravitational-wave signals, we must be careful to exclude any injections in a window contaminated by a glitch.

Figure 2 demonstrates the issue that can arise when using a contaminated signal training set. These plots show the cumulative distributions of coincident events found as a function of inverse combined false alarm rate for a small chunk of the H1L1V1 network search at Veto Level 3. The

TABLE III. Summary of ringdown-only signal training set. Injection set 1 corresponds to coverage of the ringdown template bank in (f_0, Q) space. Injection set 2 corresponds roughly to the potential $(M_f, \hat{a})$ space accessible to the ringdown-only search.

	Injection set 1	Injection set 2
Distribution:	Uniform in (f_0, Q)	Uniform in $(M_f, \hat{a})$
Parameter 1:	$f_0/\mathrm{Hz} \in [50, 2000]$	$M_f/\mathrm{M}_\odot \in [50, 900]$
Parameter 2:	$Q \in [2, 20]$	$\hat{a} \in [0, 0.99]$

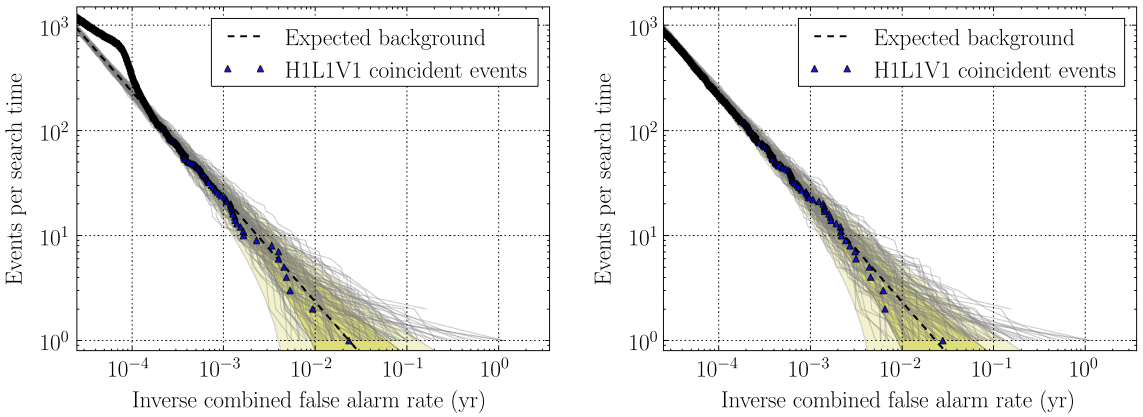


FIG. 2 (color online). Cumulative distributions of coincident events found as a function of inverse combined false alarm rate. The data plotted here show results for a ringdown-only search over ~9 days of H1L1V1 network data at Veto Level 3. Blue triangles represent the coincident events found by the ringdown-only search [7]. Grey lines plot the coincident events in each of the 100 time-shift experiments. Yellow contours mark the 1σ and 2σ regions of the expected background from accidental coincidences. (a) The results of the search obtained with a RFBDT classifier trained on a contaminated injection set. (b) The results when the classifier is trained on a clean injection set. A RFBDT classifier trained on a clean injection set properly ranks H1L1V1 coincidences with low significance so that there is not a "bump" in the distribution at low combined FAR.

ranking statistic used in both Figs. 2(a) and 2(b) is M_{forest}. However, results for Fig. 2(a) were obtained with a RFBDT classifier trained on injections identified using an injection-finding window of width ± 1.0 s (i.e., a contaminated injection set). Results for Fig. 2(b) were obtained with a RFBDT classifier trained on injections identified using a narrower injection-finding window of width ± 0.01 s and after removing any injections made within ± 0.5 s of a glitch (i.e., a clean injection set). In Fig. 2(a), we see that there is an excursion of H1L1V1 gravitational-wave candidate coincidences from the 2σ region of the expected background at low values of inverse combined FAR. This excursion for coincidences with low significance is caused by a population of injections that were misidentified because of a nearby glitch in the data. The RFBDT classifier was taught that these glitches designated as injections should be classified as signal. Thus, when similar glitches were found as coincidences in the H1L1V1 network search at Veto Level 3, they were given a boost in their M_{forest} rank. However, in Fig. 2(b), we see that by excluding these misidentified injections from the training set, the low significance H1L1V1 coincidences now fall within the 2σ region of the expected background.

We developed a software tool[3] for use with the LALSuite gravitational-wave data analysis routines[4] to construct clean injection sets. Using this tool, we investigated two time window parameters that can be tuned: the width of the injection-finding time window and the width of the injection-removal time window.

The injection-finding time window is motivated by the fact that a trigger due to an injection should be found in the data within a few milliseconds of the injection time given the light travel time between detectors. Thus, in Gaussian detector noise, a few millisecond-wide injection-finding window should be sufficient. However, because of non-Gaussian, nonstationary detector noise, the coincidence of triggers associated with an injection could be overshadowed if a loud glitch trigger is nearby. Thus, we allow a much larger window. When conducting searches for gravitational waves, this window is typically set to ± 1.0 s from the injection time. However, such a large window results in a contaminated signal training set as we see in Fig. 2.

The injection-removal time window is motivated by the fact that a significant trigger found by the search before injections are performed is a potential contaminating trigger for any injection made similarly in time. A simple time window is used to cross-check whether an injection trigger found by the search could be attributed to a trigger found in detector data before injections were performed.

We investigated separately the performance of the RFBDT classifier for the ringdown-only search for detection of $q = 1$ and $q = 4$ EOBNRv2 simulated waveforms at Veto Level 3. We performed several tuning runs, adjusting the size of the injection-finding and injection-removal windows. We found that an injection-finding window of ± 0.01 s around an injection and an injection-removal window of ± 0.5 s around an injection were the most effective at combating the excess of foreground triggers at low significance. These settings were used in

[3]https://ligo-vcs.phys.uwm.edu/cgit/lalsuite/tree/pylal/bin/ligolw_dbinjfind.

[4]https://www.lsc-group.phys.uwm.edu/daswg/projects/lalsuite.html.

designing both the IMR and ringdown-only signal training sets.

2. Background training set

The background training set composed of accidental coincidences is not noticeably contaminated by signal. Since this background is constructed by time shifting the data, it is possible that a real gravitational-wave signal could result in time-shifted triggers contaminating the background training set. However, given the rare detection rates for gravitational waves in the detector data analyzed here, it is unlikely that such a contamination has occurred. However, in the advanced detector era, when gravitational-wave detection is expected to be relatively common, this issue will need to be revisited.

An additional issue to consider for the background training set concerns the size of the set. In Refs. [6,7] and briefly in Sec. III, we describe the procedure used to compute the upper limit on BBH and IMBH coalescence rates. The typical procedure involves analyzing data in epochs of ∼1 to 2 months. For these investigations, we ran the RFBDT classifier for each ∼1 to 2 month chunk. However, the size of the background training set for these ∼1 to 2 month analyses can be as small as 1% of the total background training set available for the entire Period 2 analysis. Thus, for the ringdown-only search, we took the additional step of examining the performance of the RFBDT classifier in the case where the monthly analyses used background training sets from their respective months and in the case where the monthly analyses used background training sets from the entire Period 2 analysis. As we report in Sec. VII, using a background training set composed of time-shifted coincidences from the entire Period 2 analysis does not result in a clear sensitivity improvement.

C. Feature vector

A multivariate statistical classifier gives us the ability to use all available gravitational-wave data analysis statistics to calculate a combined FAR. These may include single trigger statistics such as SNR and the χ^2 signal consistency test [63] as well as empirically designed composite statistics that were previously used by each search as a classification statistic. The classifier will inherit the distinguishing power of the composite statistics as well as any other information we provide from statistics that may not have been directly folded into the composite statistics. These could include information that highlights inconsistencies in the single triggers' template parameters or alerts us to the presence of bad data quality. Also, although we did not include them here, coherent and null SNRs computed from coherent analyses [71,77] may be available to a search. The set of all statistics characterize the feature space and each coincidence identified by the search is described by a feature vector.

As explained in Sec. V A, a different subset of feature vector parameters is chosen at each node. Selecting the optimal size of the subset can increase the randomness of the forest and reduce concerns of overfitting. We discuss the tuning of this number in Sec. VI D 3.

Also, note that the RFBDT algorithm can only make plane cuts through the feature space. It cannot reproduce a statistic that is composed of a nonlinear combination of other statistics. As we describe below, if we know *a priori* a useful functional form for a nonlinear composite statistic, we should include that statistic in the feature vector. Such a statistic can only ever be approximated by the plane cuts. Nevertheless, we design feature vectors with a large selection of statistics in the hope that some combination may be useful.

Details of the parameters chosen to characterize coincidences with the RFBDT classifier for the IMR and ringdown-only searches are given in Secs. VI C 1 and VI C 2 and are summarized in Tables IV and V.

1. IMR search feature vector

In the following sections, we provide more details on the definitions of each statistic that defined the IMR search feature vector. Density distributions of these statistics for this search's simulated signal and background training sets are shown in Figs. 8 and 9 in Appendix A.

a. Single trigger statistics

Single trigger statistics are defined for each individual trigger that makes up each multidetector coincidence. For the IMR search, single trigger statistics added to the feature vector included the matched-filter SNR from each detector [51], the χ^2 signal-consistency test for the matched-filter result in a number of frequency bins for each detector [63], the r^2 veto duration for which a weighted χ^2 exceeds a preset threshold, and the $\chi^2_{\text{continuous}}$ quantity for the residual of the SNR and autocorrelation time series in each detector.

More details on the matched-filter SNR for the IMR search templates are given in [14] and a definition is given in Table IV.

The χ^2 signal-consistency test, only currently calculated for the low and high mass searches, tests how well the template waveform matches the data in various frequency bands. The bins are constructed so that the matched template contributes an equal amount of SNR to each bin. Then the following quantity is computed:

$$\chi^2 = 10\left[\sum_{i=0}^{10}(\rho_i - \rho/10)^2\right], \tag{10}$$

where ρ_i is the SNR contribution from the ith bin.

The r^2 veto duration measures the amount of time that the following quantity is above a threshold of 0.0002, within 6 s of the trigger:

TABLE IV. Feature vector parameters for the IMR search's RFBDT classifier. The quantities indexed by i are included for both detectors a and b.

Quantity	Definition	Description
ρ_i	$\frac{\lvert\langle x_i,h_i\rangle\rvert}{\sqrt{\langle h_i,h_i\rangle}}$	Signal-to-noise ratio of trigger found in detector i; found by filtering data $x_i(t)$ with template $h_i(t,\mathcal{M},\eta)$; described in Ref. [51]
χ^2_i	$10[\sum_{j=0}^{10}(\rho_j-\rho_i/10)^2]$	Quantity measuring how well the data match the template in j frequency bins for detector i; derived in Ref. [63]
$\rho_{\mathrm{eff},i}$	$\frac{\rho_i}{[\frac{\chi^2_i}{10}(1+\rho_i^2/50)]^{1/4}}$	Effective signal-to-noise ratio of the trigger found in detector i; used as a ranking statistic in Ref. [70]
r^2_i veto duration	$\{t\colon \frac{\chi^2_i(t)}{10+\rho_i(t)^2}>10\}$	Duration t that a weighted-χ^2 time series in detector i goes above a threshold of 10; described in Ref. [78]
$\chi^2_{\mathrm{continuous},i}$	$\sum\lvert\langle x_i(t)x_i(t+\tau)\rangle-\rho_i\rvert^2$	Sum of squares of the residual of the SNR time series and the autocorrelation time series of a single detector trigger from detector i; described in Ref. [79]
$\rho_{\mathrm{high,combined}}$	$\sqrt{\sum_i^N \rho^2_{\mathrm{high},i}}$	Combined IMR search's reweighed signal-to-noise ratio used as a ranking statistic in Ref. [6] where i sums over N triggers found in coincidence
$\mathrm{d}t$	$\lvert t_a-t_b\rvert$	Absolute value of time difference between triggers in detectors a and b
$\mathrm{d}\mathcal{M}_{\mathrm{rel}}$	$\frac{\lvert\mathcal{M}_a-\mathcal{M}_b\rvert}{\mathcal{M}_{\mathrm{average}}}$	Absolute value of the relative difference in the chirp mass of the templates matched to the data in detectors a and b
$\mathrm{d}\eta_{\mathrm{rel}}$	$\frac{\lvert\eta_a-\eta_b\rvert}{\eta_{\mathrm{average}}}$	Absolute value of the symmetric mass ratio of the templates matched to the data in detectors a and b
e-thinca	E	Value of the ellipsoidal coincidence test, which measures the distance of the two matched templates in time-mass parameter space; derived in Ref. [60]

$$r^2=\frac{\chi^2(t)}{p+\rho(t)^2}, \tag{11}$$

where $p=10$ bins are used. This quantity is motivated by the fact that a signal is unlikely to exactly match a template so a noncentrality parameter is introduced to the distribution of the χ^2 signal-consistency test. Thus, rather than thresholding on χ^2, we threshold on r^2.

The $\chi^2_{\mathrm{continuous}}$ calculation performs a sum of squares of the residual of the SNR time series and the autocorrelation time series of a single detector trigger.

b. Composite statistics

Composite statistics are defined by combining single trigger statistics in a meaningful way and are computed once for each coincidence. Although the classifier can approximate such statistics in the multidimensional parameter space (e.g., if they are a combination of the ρ and χ^2), this ability is limited by the tree depth, the number of decision tree cuts before hitting the minimum leaf size. Thus, if we have *a priori* knowledge of a useful functional form for a ranking statistic, we should provide the classifier with this information. By providing this information up front, a classifier can improve upon these good statistics rather than trying to construct them itself.

Some of these composite statistics have previously been used as ranking statistics when calculating combined FARs in searches. For the IMR search, we include several previous ranking statistics in the feature vector.

The first of these is known as the effective SNR statistic and was used as a ranking statistic in [14],

$$\rho_{\mathrm{eff}}=\frac{\rho}{[\frac{\chi^2}{10}(1+\rho^2/50)]^{1/4}}. \tag{12}$$

The second is known as $\rho_{\mathrm{high,combined}}$, a χ^2-weighted statistic described in detail in [6] for the IMR search. Because of the different distributions of background triggers over SNR and χ^2 for longer-duration versus shorter-duration templates, a different choice of ranking statistics was selected for each bin in [6]. For long duration events, the following was used:

$$\hat{\rho}=\begin{cases}\frac{\rho}{[(1+(\chi^2_r)^3)]^{1/6}} & \text{for } \chi^2_r>1\\ \rho & \text{for } \chi^2_r\le 1\end{cases}, \tag{13}$$

where $\chi^2_r\equiv\chi^2/(2p-2)$ for number of frequency intervals $p=10$. For shorter duration events, Eq. (12) was used. Thus, $\rho_{\mathrm{high,combined}}$ is a piecewise function of ρ_{eff} and $\hat{\rho}$ and is combined as a quadrature sum of single-detector statistics.

Additionally, we calculate quantities that provide an indication of how close the pair of triggers from different detectors are in the metric space $(\mathcal{M},\eta,t)$ for the IMR search. These include the difference in arrival time $\mathrm{d}t$, the relative difference in chirp mass $\mathrm{d}\mathcal{M}_{\mathrm{rel}}$, the relative difference in the symmetric mass ratio $\mathrm{d}\eta_{\mathrm{del}}$, and a quantity known as the *e-thinca* test that combines these three by constructing error ellipsoids in time and mass space [60].

2. Ringdown-only feature vector

In the following sections, we provide more details on the definitions of each statistic that defined the ringdown-only

TABLE V. Feature vector parameters for the ringdown-only search's RFBDT classifier. The quantities indexed by i are included for both detectors a and b.

Quantity	Definition	Description
ρ_i	$\frac{\lvert\langle x_i,h_i\rangle\rvert}{\sqrt{\langle h_i,h_i\rangle}}$	Signal-to-noise ratio of trigger found in detector i; found by filtering data $x_i(t)$ with template $h(t,f_i,Q_i)$
$\mathrm{d}t$	$\lvert t_a - t_b\rvert$	Absolute value of time difference between triggers in detectors a and b
$\mathrm{d}f$	$\lvert f_a - f_b\rvert$	Absolute value of template frequency difference between triggers in detectors a and b
$\mathrm{d}Q$	$\lvert Q_a - Q_b\rvert$	Absolute value of template quality factor difference between triggers in detectors a and b
$\mathrm{d}s^2$	$g_{ij}\mathrm{d}p^i\mathrm{d}p^j$	Three-dimensional metric distance between two triggers in (f_0,Q,t) space for $p\in(f_0,Q,t)$; outlined in Ref. [7]
g_{tt}	$\pi^2 f_0^2\frac{1+4Q^2}{Q^2}$	Metric coefficient in (t,t) space
$g_{f_0f_0}$	$\frac{1+6Q^2+16Q^4}{4f_0^2(1+2Q^2)}$	Metric coefficient in (f_0,f_0) space
g_{QQ}	$\frac{1+28Q^4+128Q^6+64Q^8}{4Q^2(1+6Q^2+8Q^4)}$	Metric coefficient in (Q,Q) space
g_{tf_0}	$2\pi Q\frac{1+4Q^2}{1+2Q^2}$	Metric coefficient in (t,f_0) space
g_{tQ}	$2\pi f_0\frac{1-2Q^2}{(1+2Q^2)^2}$	Metric coefficient in (t,Q) space
g_{f_0Q}	$\frac{1+2Q^2+8Q^4}{2f_0Q(1+2Q^2)^2}$	Metric coefficient in (f_0,Q) space
ξ	$\max(\frac{\rho_a}{\rho_b},\frac{\rho_b}{\rho_a})$	Maximum of the ratio of signal-to-noise ratios for triggers a to b or b to a
${\rho_N}^2$	$\sum_i^N {\rho_i}^2$	Combined network signal-to-noise ratio for N triggers found in coincidence
ρ_{S4}	$\rho_{\mathrm{S4,triple}}$, $\rho_{\mathrm{S4,double}}$	Detection statistic used in Ref. [47]; outlined in Eqs. (15) and (16)
$\rho_{\mathrm{S5/S6}}$	$\rho_{\mathrm{S5/S6,triple}}$, $\rho_{\mathrm{S4,double}}$	Detection statistic described in Ref. [71]; outlined in Eqs. (17) and (16)
D_i	$\frac{\sigma_i}{\rho_i}(1\ \mathrm{Mpc})$	Effective distance of trigger found with signal-to-noise ratio ρ_i in detector i that has a sensitivity σ_i to a signal at 1 Mpc
$\mathrm{d}D$	$\lvert D_a - D_b\rvert$	Absolute value of effective distance difference between triggers in detectors a and b
κ	$\max(\frac{D_a}{D_b},\frac{D_b}{D_a})$	Maximum of the ratio of effective distances for triggers a to b or b to a
n_i	$n_i(\lvert t\rvert < 0.5\ \mathrm{ms})$	Count of the number of triggers in detector i clustered over a time interval of 0.5 ms using the SNR peak-finding algorithm in Ref. [80]
hveto_i	$\begin{cases}1: & hveto\ \text{flag on}\\ 0: & hveto\ \text{flag off}\end{cases}$	Binary value used to indicate whether a trigger in detector i occurred during a *hveto* time interval [81] flagged for noise transients

search feature vector. Density distributions of these statistics for this search's simulated signal and background training sets are shown in Figs. 10–13 in Appendix A.

a. Single trigger statistics

For the ringdown-only search, single trigger statistics added to the feature vector included the matched-filter SNR from each detector and the effective distance from each detector, $\mathrm{D_{eff}}$.

More details on the matched-filter SNR, specifically for ringdown templates, are given in [7,47].

The effective distance is equivalent to the distance r to a source that is optimally oriented and located. The theoretical formula for the effective distance is defined in terms of the F_+ and $F_\times$ detector antenna pattern functions and the inclination angle ι,

$$D_{\mathrm{eff}} = \frac{r}{\sqrt{F_+^2(1+\cos^2\iota)/4 + F_\times^2\cos^2\iota}}. \tag{14}$$

In practice, however, the effective distance is calculated from the power spectral density of the detector and the matched-filter SNR; see Table V.

b. Composite statistics

Composite statistics included in the feature vector for the ringdown-only search include a combined network SNR, a detection statistic used in [47], and a ranking statistic detailed in [71,82,83].

The combined network SNR for the N detectors participating in the coincidence,

$$\rho_N^2 = \sum_i^N \rho_i^2, \tag{15}$$

where ρ_i is the SNR in the ith detector, is the optimal ranking statistic for a signal with known parameters in Gaussian noise.

In the ringdown-only search in [47], because of the dearth of false alarms found in triple coincidence, a suitable statistic for ranking triple coincident events was found to be the network SNR in Eq. (15) such that $\rho_{\mathrm{S4,triple}} = \rho_N^2$. However, Ref. [47] found a high level of double coincident false alarms, often with very high SNR in only one detector. While it is possible that a real gravitational-wave source could have an orientation that would produce an asymmetric SNR pair, the occurrence is relatively rare in comparison to the occurrence of this feature for false alarms. The network SNR is clearly nonoptimal in this case. References [80,84] found the optimal statistic in such a case to be a "chopped-L" statistic,

$$\rho_{\mathrm{S4,double}} = \min\left\{\begin{array}{c}\rho_{\mathrm{ifo1}} + \rho_{\mathrm{ifo2}} \\ \alpha\rho_{\mathrm{ifo1}} + \beta \\ \alpha\rho_{\mathrm{ifo2}} + \beta\end{array}\right\}, \tag{16}$$

where the tunable parameters α and β were set to 2 and 2.2, respectively, as described in [47]. We include this piecewise detection statistic composed of $\rho_{\mathrm{S4,triple}}$ and $\rho_{\mathrm{S4,double}}$ in the feature vector.

For the most recent ringdown-only search [7], because of a large increase in analysis time and lower SNR thresholds, a significant population of triple coincident false alarms was found. Thus, an additional chopped-L–like statistic was developed for triple coincidences,

$$\rho_{\mathrm{S5/S6,triple}} = \min\left\{\begin{array}{c}\rho_N \\ \rho_{\mathrm{ifo1}} + \rho_{\mathrm{ifo2}} + \gamma \\ \rho_{\mathrm{ifo2}} + \rho_{\mathrm{ifo3}} + \gamma \\ \rho_{\mathrm{ifo3}} + \rho_{\mathrm{ifo1}} + \gamma\end{array}\right\}, \tag{17}$$

where the tunable parameter γ was set to 0.75. The development and tuning of this new statistic are described in detail in [71,82,83]. Again, we include this piecewise detection statistic composed of $\rho_{\mathrm{S5/S6,triple}}$ and $\rho_{\mathrm{S4,double}}$ in the feature vector.

In addition to these three previous ranking statistics, we include the following simple composite statistics: the maximum of the ratios of the SNRs for triggers in each detector, the difference in recovered effective distances, and the maximum of the ratios of the recovered effective distances.

Finally, we calculate quantities that provide an indication of how close the pair of triggers from different detectors are in the metric space (f_0, Q, t) for the ringdown-only search. These include the difference in arrival time $\mathrm{d}t$, the template frequency difference $\mathrm{d}f_0$, the template quality factor difference $\mathrm{d}Q$, and the three-dimensional metric distance $\mathrm{d}s^2$ between two triggers in (f_0, Q, t) space [7,61]. Also included are the three-dimensional coincidence metric coefficients g_{tt}, $g_{f_0 f_0}$, g_{QQ}, g_{tf_0}, g_{tQ}, and $g_{f_0 Q}$ defined in Table V.

c. Other parameters

Two additional parameters were added to the feature vector for the ringdown-only search in an effort to provide data quality information to the classifier.

The first was a binary value used to indicate whether a trigger in a coincidence occurred during a time interval flagged for noise transients. The flagged intervals were defined using the hierarchical method for vetoing noise transients known as *hveto* as described in [81]. The LIGO and Virgo gravitational-wave detectors have hundreds of auxiliary channels monitoring local environment and detector subsystems. The hveto algorithm identifies auxiliary channels that exhibit a significant correlation with transient noise present in the gravitational-wave channel and that have a negligible sensitivity to gravitational waves. If a trigger in the gravitational-wave channel is found to have a statistical relationship with auxiliary channel glitches, a flagged time interval is defined. In Sec. VII C, we explore the performance of the RFBDT classifier before and after the addition of the *hveto* parameter to the feature vector. This investigation was done to explore the ability of the classifier to incorporate data quality information.

The second additional parameter was a count of the number of single detector triggers clustered over a time interval of 0.5ms using a SNR peak-finding algorithm described in [80]. The motivation behind this parameter comes from investigations that show that a glitch will be recovered with a different pattern of templates over time than a signal [79]. Ideally, a χ^2-based statistic could be computed. However, in the absence of this test for the ringdown-only search, we simply provide a count of the number of templates in a small time window around each trigger giving a matched-filter SNR above the threshold.

Significant work has been done to identify glitches in the data using multivariate statistical classifiers [32] and Bayesian inference [85]. With more development, this work could be used to provide information to a multivariate classifier used to identify gravitational waves, allowing for powerful background identification and potentially significant improvement to the sensitivity of the search.

D. Random forest parameters

A summary of the tunable parameters selected for the RFBDT algorithm for each search is given in Table I.

1. Number of trees

We can adjust the number of trees in our forest to provide a more stable M_{forest} statistic. Increasing the number of trees results in an increased number of training events folded into the M_{forest} statistic calculation. However, the training data contain a finite amount of information, and adding a large number of additional trees will ultimately reproduce results found in earlier trees. Furthermore, adding more trees will increase the computational cost of training linearly.

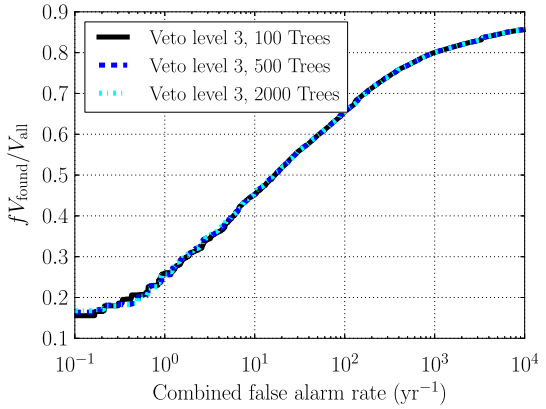


FIG. 3 (color online). Investigation of the effect of using a different number of trees on the recovery of $q = 1$ EOBNRv2 simulated waveforms at Veto Level 3. In general, we find that the use of more than 100 trees gives roughly the same sensitivity regardless of mass ratio or veto level. In this ROC, to adjust for the loss in analysis time in moving from Veto Level 2 to Veto Level 3, we scale the volume fraction in Eq. (8) by the ratio of analysis times $f = t_{\rm VL3}/t_{\rm VL2}$. From the analysis times reported in Sec. III, we find $f = 0.88$.

In Fig. 3, we investigate the effect of using a different number of trees for the ringdown-only search on the recovery of $q = 1$ EOBNRv2 waveforms at Veto Level 3. We find no significant improvement for using more than 100 trees. Similar results were obtained at Veto Level 2 and for the recovery of $q = 4$ EOBNRv2 waveforms. The IMR search trained classifiers with 100 trees in each forest. Initially, for the ringdown-only search, we selected to use 2000 trees in order to offset possible loss in sensitivity due to needing a larger leaf size as described in Sec. VI D 2. However, we ultimately found that this did not change the sensitivity. Since computational costs were not high, we left the forest size as 2000 trees for the ringdown-only search.

2. Minimum leaf size

The minimum leaf size defines the stopping point for the splitting of nodes. We define the minimum number of events allowed in a node before it becomes a leaf. When all nodes become leaves, the recursive splitting of the tree stops.

The choice of leaf size is affected by how representative the training data are of actual data and by how many coincident events are in the training data. If the leaf size is too small, the forest will be overtrained. In this case the sort is specific to the training data and may be systematically wrong for anything else. If the leaf size is too large, the forest will be undertrained. The individual decision trees did not have enough chances to make good cuts on the training data. We will be left unsure if any coincident event is signal or background.

We are limited by the size of the background training set of time-shifted data. In each monthly analysis, the size of the background training set varied between thousands to hundreds of thousands of coincident events depending on veto level and analyzed networks. A leaf size of 5 worked very well for the IMR search's trees, but investigations on the ringdown-only search with a leaf size of 5 showed that such a small choice led to an overtrained forest. Some signal *and* background coincidences were given an infinite $M_{\rm forest}$ rank (i.e., the classifier was 100% sure that the coincidence was signal). By exploring leaf sizes around 0.1%–1% the size of the varied background training sets, we found that a leaf size of 65 eliminated the overtraining and also gave good performance for the ringdown-only search.

3. Number of sampled parameters

At each node, we choose a random subset of parameters to use for splitting. Out of N_v total feature vector parameters, we select m randomly and evaluate the split criteria for each. Thus, a different set of m parameters is available for picking the optimal parameter and its threshold for each branching point.

If m is too large, each node will have the same parameters available to make the splits. This can lead to the same few parameters being used over and over again, and the forest will not fully explore the space of possible cuts. Furthermore, because individual trees will be making cuts based on the same parameters, all of the trees in the forest will be very similar. This is an example of overtraining.

If m is too small, each node would have very few options to make the splits. The classifier would be forced to use poor parameters at some splits, resulting in inefficient cuts. The tree can run up against the leaf size limit before the training events were well sorted. This is an example of undertraining. The classification in this case would be highly dependent on the presence (or lack thereof) of poor parameters.

A general rule of thumb for a good number of random sampled parameters is $\sim\sqrt{N_v}$ [86]. For the IMR search, of the 15 parameters that make up the feature vector, we empirically found good performance for a selection of 6 randomly chosen parameters at each node. For the ringdown-only search, 14 out of the total 24 feature vector parameters gave good performance.

4. Criterion for optimization

The optimization criterion is used to select the best thresholds on parameters and proceeds the selection of random sampled parameters for each node. The RFBDT algorithm provides several methods to determine the optimal parameter thresholds. These are grouped by whether the output is composed of a discrete set or a continuous set of $M_{\rm forest}$ rankings. While some of the discrete statistics

performed well, we preferred to draw rankings from a continuous set. Of the optimization criteria that gave continuous statistics, the Gini index [76] and negative cross entropy (defined in Sec. V A) gave good performance and were comparable to each other for both searches. Additionally, in order to obtain a good average separation between signal and background, the suggested optimization criteria are either Gini index or negative cross entropy [74]. Thus, these two statistics were chosen for the IMR and ringdown-only searches, respectively. The choices were arbitrary in the sense that either optimization criteria would have been suitable for either search. Splits were made only if they minimized the Gini index or the negative cross entropy.

VII. RESULTS

A. IMR search

In order to assess the sensitivity improvements of the IMR search to waveforms from BBH coalescing systems with nonspinning components, we use the same set of EOBNRv2 injections used to compute the upper limits on BBH coalescence rates in Sec. V B of [6]. These injections were distributed approximately uniformly over the component masses m_1 and m_2 within the ranges $1 \leq m_i/\mathrm{M}_\odot \leq 99$ and $20 \leq M/\mathrm{M}_\odot \leq 109$. Additionally, we use the same set of IMRPhenomB injections used to make statements on sensitivity to spinning and nonspinning BBH coalescences in Sec. V C of [6]. We use a nonspinning set and a spinning set of IMRPhenomB injections, both uniformly distributed in total mass $25 \leq M/\mathrm{M}_\odot \leq 100$ and uniformly distributed in $q/(q+1) = m_1/M$ for a given M, between the limits $1 \leq q < 4$. In addition, the spinning injections were assigned (anti)aligned spin parameter χ_s uniformly distributed between -0.85 and 0.85.

The previous IMR search over Period 2 data [6] used the combined signal-to-noise and χ^2-based ranking statistic $\rho_{\mathrm{high,combined}}$ for FAR calculations. For more details on $\rho_{\mathrm{high,combined}}$, see Sec. VI C 1 and Table IV. Here, we report on a reanalysis that replaces $\rho_{\mathrm{high,combined}}$ with the ranking statistic calculated by the RFBDT, M_{forest}, as described in Sec. V A. Additionally, we have chosen a different FAR threshold for calculating sensitivity, rather than the loudest event statistic typically used in calculating upper limits in [6]. The threshold that we use is the expected loudest FAR,

$$F\check{A}R = 1/T, \tag{18}$$

where T is the total time of the analysis chunk being considered. For a listing of $F\check{A}R$ for each analysis chunk and a comparison with the loudest event statistic, see Table 8.1 of [17].

Improvements in the following section are reported with uncertainties determined using the statistical uncertainty originating from the finite number of injections that we have performed in these investigations.

B. IMR search sensitive VT improvements

Figure 4 demonstrates the percent improvements in sensitive volume multiplied by analysis time (VT) when using the M_{forest} ranking statistic, rather than the $\rho_{\mathrm{high,combined}}$ ranking statistic. Results are shown at both Veto Levels 2 and 3 for total binary masses from $25 \leq M/\mathrm{M}_\odot \leq 100$ in mass bins of width 12.5 $\mathrm{M}_\odot$. Improvements for EOBNRv2 waveforms are shown in Fig. 4(a) and for IMRPhenomB are shown in Fig. 4(b). The use of the M_{forest} ranking statistic gives improvements in VT over the use of $\rho_{\mathrm{high,combined}}$ at both Veto Levels 2 and 3.

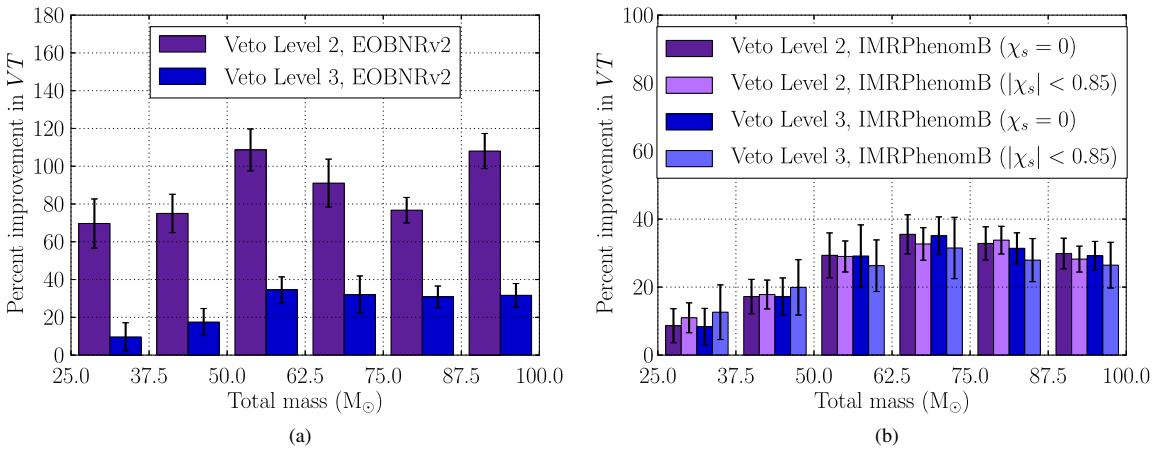


FIG. 4 (color online). Percent improvements (over the use of the $\rho_{\mathrm{high,combined}}$ [6] ranking statistic at Veto Levels 2 and 3) in sensitive volume multiplied by analysis time (VT) for the recovery of EOBNRv2 (a) simulated waveforms at Veto Levels 2 and 3 and the recovery of IMRPhenomB (b) waveforms at Veto Levels 2 and 3 by the IMR search. Results for IMRPhenomB are shown for signals with spin parameter ($|\chi_s| < 0.85$) and no spin ($\chi_s = 0$). The quantity VT gives us a measure of the true sensitivity of the search and allows us to compare performances across veto levels. Results are shown for total binary masses from $25 \leq M/\mathrm{M}_\odot \leq 100$ in mass bins of width 12.5 $\mathrm{M}_\odot$.

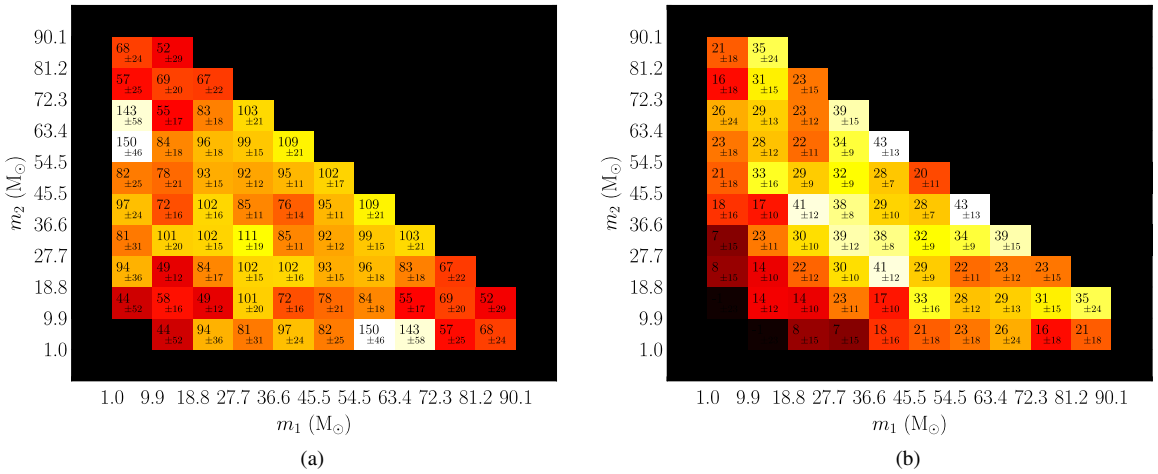


FIG. 5 (color online). Percent improvements (over the use of the $\rho_{\text{high,combined}}$ [6] ranking statistic at Veto Levels 2 and 3) in sensitive volume multiplied by analysis time (VT) for the recovery of EOBNRv2 simulated waveforms at Veto Levels 2 (a) and 3 (b) by the IMR search. Percent improvement results are shown as a function of binary component masses. Note that the color scales of (a) and (b) are not equivalent.

The largest improvements are seen for total masses larger than 50 $M_\odot$. The IMR search is more sensitive in these higher mass regions. Thus, larger improvement is found where the search is more sensitive.

For EOBNRv2 waveforms, larger improvements are seen at Veto Level 2 than at Veto Level 3. At Veto Level 2, VT improvements ranged from $70_{\pm 13}\%$ to $109_{\pm 11}\%$ for EOBNRv2 waveforms and from roughly $9_{\pm 5}\%$ to $36_{\pm 6}\%$ for IMRPhenomB waveforms. At Veto Level 3, VT improvements ranged from $10_{\pm 8}\%$ to $35_{\pm 7}\%$ for EOBNRv2 waveforms and remained roughly the same for IMRPhenomB waveforms. More investigation is needed to understand why IMRPhenomB improvements are not as strong as EOBNRv2 improvements. One contributing factor could be component spin, which introduces several competing effects on the search including increased horizon distance with positive χ_s, decreased sensitivity due to reduced overlap with EOBNRv1 templates, and higher signal-based χ^2 test values [6]. It is currently unclear if any of these effects reduce the potential percent improvement seen with the M_{forest} ranking statistic.

For more detail, Fig. 5 shows the percent improvements in VT for EOBNRv2 waveforms as a function of component masses. At Veto Level 2 in Fig. 4(a), we see that every mass bin sees a percent improvement in VT. At Veto Level 3 in Fig. 4(b), again we see that the improvements are smaller than at Veto Level 2. In fact, no improvement is found for the lowest mass bin centered on $(5.5, 14.4)$ $M_\odot$.

In Table VI, we explore the percent VT improvements obtained with M_{forest} at different veto levels. The improvements reported are made with respect to the sensitive volumes achieved with the $\rho_{\text{high,combined}}$ ranking statistic at Veto Level 2. These values are presented as a means of comparing sensitivity between Veto Level 2 and Veto Level 3. We see that M_{forest} at Veto Level 2 shows greater improvement and hence a more stringent upper limit than M_{forest} at Veto Level 3. This is in contrast to the better performance of $\rho_{\text{high,combined}}$ at Veto Level 3 than at Veto Level 2. For the standard IMR search with the $\rho_{\text{high,combined}}$ ranking statistic, the additional vetoing of poor quality data at Veto Level 3 was performed with the goal of preventing high SNR noise events from contaminating the list of gravitational-wave candidate events and reducing the sensitivity of the search. However, for the random forest technique, those high SNR noise events are down weighted in significance due to information contained in other parameters in the feature vector. As a search at Veto Level 2 has more analysis time, it has the potential to have better sensitivity than a search at Veto Level 3. In Table VI, we see that the use of the M_{forest} ranking statistic for the IMR search has resulted in a better search sensitivity

TABLE VI. Percent VT improvements over the use of the $\rho_{\text{high,combined}}$ ranking statistic in the IMR search at Veto Level 2 for EOBNRv2 waveforms. Note that these percent improvements should not be compared with values reported in Figs. 4 and 5 but are rather presented as a means of comparing sensitivity between Veto Level 2 and Veto Level 3. We see that M_{forest} at Veto Level 2 shows greater improvement and hence a more stringent upper limit than M_{forest} at Veto Level 3.

Mass bin ($M_\odot$)	$\rho_{\text{high,combined}}$ Veto Level 3	M_{forest} Veto Level 2	M_{forest} Veto Level 3
25.0–37.5	$28\% \pm 8\%$	$70\% \pm 13\%$	$40\% \pm 9\%$
37.5–50.0	$36\% \pm 8\%$	$75\% \pm 10\%$	$60\% \pm 10\%$
50.0–62.5	$35\% \pm 7\%$	$109\% \pm 11\%$	$82\% \pm 10\%$
62.5–75.0	$33\% \pm 10\%$	$91\% \pm 13\%$	$76\% \pm 13\%$
75.0–87.5	$15\% \pm 5\%$	$77\% \pm 7\%$	$50\% \pm 6\%$
87.5–100.0	$41\% \pm 7\%$	$108\% \pm 9\%$	$86\% \pm 9\%$

at Veto Level 2. As we discuss in Sec. VII C, the ringdown-only search did not see the same behavior at Veto Level 2. The information contained in the ringdown-only search's feature vector may not have had sufficient signal and background separation information to overcome the level of background contamination present at Veto Level 2 as compared to Veto Level 3.

C. Ringdown-only search

In order to assess the sensitivity improvements of the ringdown-only search to waveforms from binary IMBH coalescing systems with nonspinning components, we use the same set of EOBNRv2 injections used to compute the upper limits on IMBH coalescence rates in Sec. V of [7]. Because of the variation in ringdown-only search sensitivity over different mass ratios, we chose to explore sensitivity improvements separately for $q = 1$ and $q = 4$. This variation occurs because the total ringdown efficiency depends on the symmetric mass ratio so that extreme mass ratio systems will not be detectable unless the system is sufficiently close [7]. The injection sets were distributed uniformly over a total binary mass range from $50 \leq M/\mathrm{M}_{\odot} \leq 450$, and upper limits were computed in mass bins of width 50 $\mathrm{M}_{\odot}$. The final black hole spins of these injections can be determined from the mass ratios and zero initial component spins [87]. For $q = 1$, we find $\hat{a} = 0.69$, and for $q = 4$, we find $\hat{a} = 0.47$.

Previous investigations of ranking statistics for the ringdown-only search [71,82,83] found that $\rho_{\mathrm{S5/S6}}$ provided better sensitivity than the ρ_{S4} ranking statistic used as a detection statistic in [47]. Thus, here we report on sensitivities based on combined FARs computed using $\rho_{\mathrm{S5/S6}}$ as a ranking statistic and using M_{forest} as a ranking statistic. We follow the same loudest event statistic procedure used in [7] for calculating upper limits. Improvements in the following section are reported with uncertainties determined using the statistical uncertainty originating from the finite number of injections that we have performed in these investigations.

Our complete investigations involve evaluating the performance of the RFBDT classifier for ringdown-only searches over Period 2 data using five separate ranking statistics, described below. Additionally, we explore the improvement separately for recovery of $q = 1$ and $q = 4$ EOBNRv2 simulated waveforms as well as for Veto Level 2 and Veto Level 3 searches.

The first ringdown-only search, to which we will compare performance, utilized the SNR-based statistic $\rho_{\mathrm{S5/S6}}$ to rank both double and triple coincident events. Details of this ranking statistic are given in Sec. VI C 2 and in [71,82,83]. In each of the investigative runs that follow, this statistic becomes a parameter that is added to the feature vector of each coincident event. A summary of the runs is given in Table VII.

TABLE VII. Summary of ringdown-only search investigations.

Run	Full background training set	*hveto* parameter
1	No	No
2	Yes	No
3	No	Yes
4	Yes	Yes

Run 1 uses RFBDTs with 2000 trees, a leaf size of 65, and a random selection of 14 parameters out of the 24 total parameters listed in Sec. VI C except the *hveto* parameter. The training set was composed of a clean signal set as outlined in Sec. VI B 1 and a background set trained separately for each ∼1–2 month chunk of Period 2 as outlined in Sec. VI B 2.

Run 2 is identical to Run 1 except that the background training set of the RFBDTs is composed of all Period 2 background coincident events rather than each corresponding ∼1–2 month set of background coincident events. We say that the RFBDTs is trained on the "full background set."

Run 3 is identical to Run 1 except that the *hveto* parameter is included in the feature vector of each coincident event. This investigation was done to explore the ability of the RFBDT to incorporate data quality information.

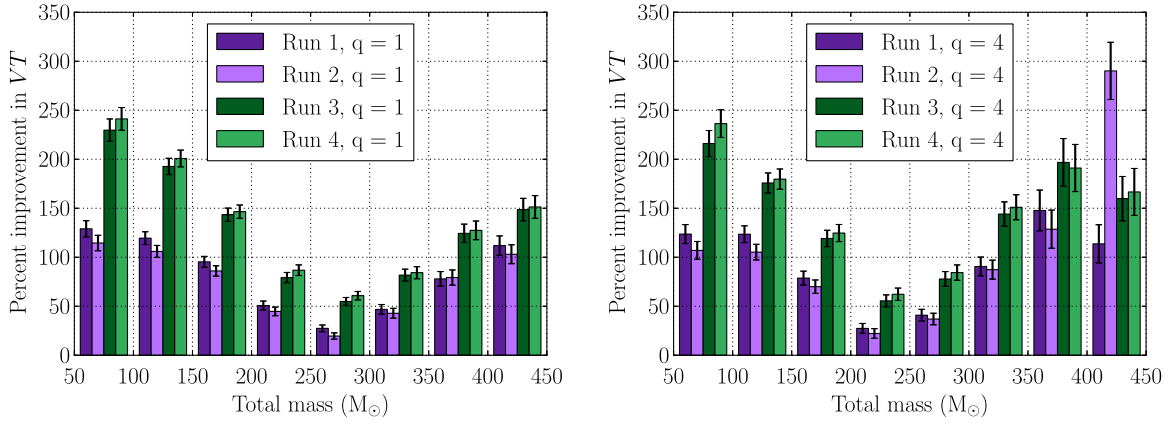
Run 4 combines the exceptions of Run 2 and Run 3. Thus, this investigation includes a RFBDT classifier trained on the full background set and feature vectors that include the *hveto* parameter.

D. Ringdown-only sensitive VT improvements

Figures 6 and 7 demonstrate the percent improvements in sensitive volume multiplied by analysis time (VT) when using the M_{forest} ranking statistic, rather than the $\rho_{\mathrm{S5/S6}}$ ranking statistic at Veto Levels 2 and 3, respectively.

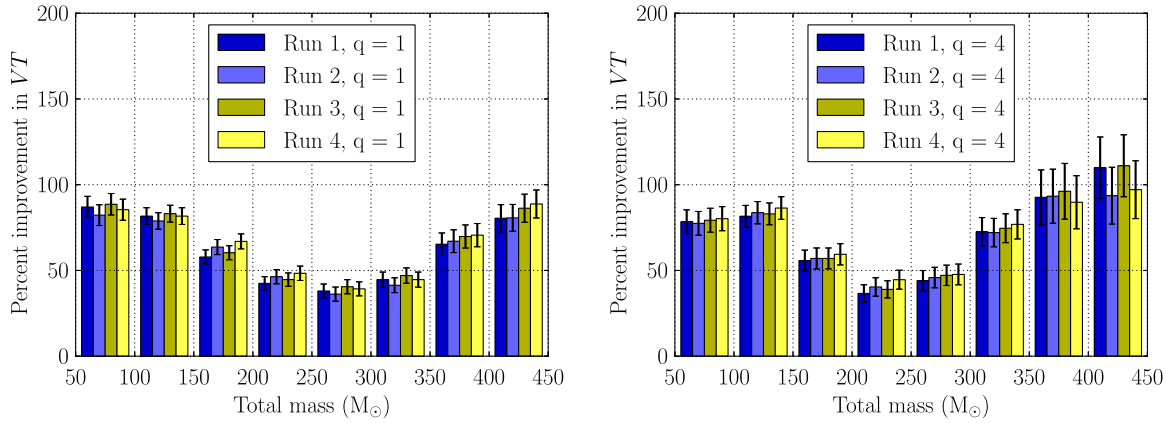
Figure 6 focuses on the comparison of Runs 1–4 over $\rho_{\mathrm{S5/S6}}$ at Veto Level 2 for each mass ratio. Here we see that all runs perform better than $\rho_{\mathrm{S5/S6}}$ at Veto Level 2. The largest percent improvements are seen in the lowest and highest mass bins. These are the mass regions where the ringdown-only search is least sensitive. Thus, in these regimes, small changes in VT lead to large percent improvements. This is the reason for the seemingly large percent improvement in Fig. 6(b) for Run 2. In general, Runs 3 and 4 that include the *hveto* parameter in the feature vector outperform Runs 1 and 2 that do not include the *hveto* parameter. Run 4 most consistently shows the largest VT improvements although the differences are not large at Veto Level 3. At Veto Level 2, VT improvements ranged from $61_{\pm 4}\%$ to $241_{\pm 12}\%$ for $q = 1$ and from $62_{\pm 6}\%$ to $236_{\pm 14}\%$ for $q = 4$.

We also note in Fig. 6 that Run 2 is slightly worse than Run 1. This is due to the fact that, generally, it is advantageous to break large analyses up into several



(a) $q = 1$ (b) $q = 4$

FIG. 6 (color online). Percent improvements (over the use of the $\rho_{S5/S6}$ ranking statistic [71] at Veto Level 2) in sensitive volume multiplied by analysis time (VT) for the recovery of $q = 1$ (a) and $q = 4$ (b) EOBNRv2 simulated waveforms at Veto Level 2. The quantity VT gives us a measure of the true sensitivity of the search and allows us to compare performances across veto levels. Results are shown for total binary masses from $50 \leq M/\mathrm{M}_\odot \leq 450$ in mass bins of width 50 $\mathrm{M}_\odot$.



(a) $q = 1$ (b) $q = 4$

FIG. 7 (color online). Percent improvements (over the use of the $\rho_{S5/S6}$ ranking statistic [71] at Veto Level 3) in sensitive volume multiplied by analysis time (VT) for the recovery of $q = 1$ (a) and $q = 4$ (b) EOBNRv2 simulated waveforms at Veto Level 3. The quantity VT gives us a measure of the true sensitivity of the search and allows us to compare performances across veto levels. Results are shown for total binary masses from $50 \leq M/\mathrm{M}_\odot \leq 450$ in mass bins of width 50 $\mathrm{M}_\odot$.

smaller chunks to account for sensitivity changes over the run. By training the RFBDTs on the full background set, we subjected the entire training set to background triggers from the least sensitive times (i.e., times when the background triggers most resembled the signal) which resulted in an overall decrease in sensitive volume. In Run 1, these troublesome background triggers would be isolated in the separate training sets for each ∼1–2 month chunk of Period 2. However, note that training the RFBDTs on the full background set with an *hveto* data quality parameter in the feature vectors results in Run 4 being more sensitive than Run 3.

Figure 7 focuses on the comparison of Runs 1–4 over $\rho_{S5/S6}$ at Veto Level 3 for each mass ratio. Again we see that all runs perform better than $\rho_{S5/S6}$ at Veto Level 3 (although percent improvements are not as large as those seen at Veto Level 2), and the largest percent improvements are seen in the lowest and highest mass bins. However, at Veto Level 3, we find that the addition of the *hveto* data quality parameter in the feature vectors of Runs 3 and 4 do not give significant improvements over Runs 1 and 2. This fact indicates that the *hveto* parameter provides no additional information on the most significant glitches for the ringdown-only search that is not already included in the vetoes at Veto Level 3. Although the difference is not large, in general, Runs 3 and 4 still outperform Runs 1 and 2. At Veto Level 3, VT improvements ranged from $39_{\pm 4}\%$ to $89_{\pm 8}\%$ for $q = 1$ and from $39_{\pm 5}\%$ to $111_{\pm 18}\%$ for $q = 4$.

TABLE VIII. Percent VT improvements over the use of the $\rho_{S5/S6}$ ranking statistic at Veto Level 2 for $q = 1$ EOBNRv2 waveforms. Note that these percent improvements should not be compared with values reported in Figs. 6 and 7 but are rather presented as a means of comparing sensitivity between Veto Level 2 and Veto Level 3. We see that Run 1 at Veto Level 3 shows greater improvement and hence a more stringent upper limit than Run 1 at Veto Level 2, unlike the IMR search as shown in Table VI. Additionally, we see that Run 1 at Veto Level 3 gives consistently larger percent VT improvements than Run 3 at Veto Level 2, indicating that the *hveto* parameter in the feature vector does not give the same sensitivity improvements as the application of traditional vetoes.

Mass bin ($M_\odot$)	$\rho_{S5/S6}$ Veto Level 3	Run 1 Veto Level 2	Run 1 Veto Level 3	Run 3 Veto Level 2	Run 3 Veto Level 3
50.0–100.0	$113\% \pm 7\%$	$129\% \pm 8\%$	$299\% \pm 14\%$	$230\% \pm 11\%$	$302\% \pm 14\%$
100.0–150.0	$92\% \pm 6\%$	$120\% \pm 7\%$	$250\% \pm 10\%$	$193\% \pm 8\%$	$252\% \pm 10\%$
150.0–200.0	$88\% \pm 5\%$	$95\% \pm 6\%$	$196\% \pm 8\%$	$143\% \pm 7\%$	$201\% \pm 8\%$
200.0–250.0	$63\% \pm 5\%$	$51\% \pm 5\%$	$133\% \pm 7\%$	$79\% \pm 5\%$	$136\% \pm 7\%$
250.0–300.0	$60\% \pm 4\%$	$27\% \pm 3\%$	$121\% \pm 7\%$	$55\% \pm 4\%$	$125\% \pm 7\%$
300.0–350.0	$66\% \pm 6\%$	$47\% \pm 5\%$	$139\% \pm 8\%$	$82\% \pm 6\%$	$143\% \pm 8\%$
350.0–400.0	$73\% \pm 8\%$	$78\% \pm 7\%$	$186\% \pm 12\%$	$124\% \pm 9\%$	$194\% \pm 12\%$
400.0–450.0	$70\% \pm 8\%$	$112\% \pm 10\%$	$207\% \pm 14\%$	$149\% \pm 11\%$	$218\% \pm 14\%$

In Table VIII, we explore the percent VT improvements obtained with M_{forest} at different veto levels with and without the *hveto* parameter. The improvements reported are made with respect to the $\rho_{S5/S6}$ ranking statistic at Veto Level 2. These values are presented as a means of comparing sensitivity between Veto Level 2 and Veto Level 3. Here we make several observations. First, unlike the behavior we observed for the IMR search, we see that Run 1 at Veto Level 3 shows greater improvement and hence a more stringent upper limit than Run 1 at Veto Level 2. Thus, the removal of poor data quality at Veto Level 3 is an important step for improving the sensitivity of the ringdown-only search. Second, we can compare Run 1 at Veto Level 3 with Run 3 at Veto Level 2. This comparison allows us to gauge whether the gain in analysis time we get by including *hveto* data quality information in the feature vector at Veto Level 2 outweighs the boost in sensitive volume we gain by removing data flagged by Veto Level 3. We see that Run 1 at Veto Level 3 gives consistently larger percent VT improvements than Run 3 at Veto Level 2. Thus, adding the *hveto* data quality information in the feature vector does not match the sensitivity improvements from the application of data quality vetoes. However, we note that the *hveto* data quality information was not specifically tuned for the ringdown-only search nor is it meant to be an exhaustive data quality investigation. Further exploration with more sophisticated data quality information is needed in order to determine whether the classifier can incorporate data quality information and approach the sensitivity achieved by the use of data quality vetoes.

VIII. SUMMARY

This paper presents the development and sensitivity improvements of a multivariate analysis applied to matched filter searches for gravitational waves produced by coalescing black hole binaries with total masses $\gtrsim 25\ M_\odot$. We focus on the applications to the IMR search which looks for gravitational waves from the inspiral, merger, and ringdown of BBHs with total mass between $25\ M_\odot$ and $100\ M_\odot$ and to the ringdown-only search which looks for gravitational waves from the resultant perturbed IMBH with mass roughly between $10\ M_\odot$ and $600\ M_\odot$. These investigations were performed over data collected by LIGO and Virgo between 2009 and 2010 so that comparisons can be made with previous IMR and ringdown-only search results [6,7]. We discuss several issues related to tuning RFBDT multivariate classifiers in matched-filter IMR and ringdown-only searches. We determine the sensitivity improvements achieved through the use of a RFBDT-derived ranking statistic over empirical SNR-based ranking statistics while considering the application of data quality vetoes. Additionally, we present results for several modifications on the basic RFBDT implementation including the use of an expansive training set and data quality information.

When optimizing the performance of RFBDT classifiers, we found that a RFBDT classifier with 100 trees, a leaf size of 5, and 6 randomly sampled parameters from the feature vector gave good performance for the IMR search while a RFBDT classifier with 2000 trees, a leaf size of 65, and 14 randomly sampled parameters from the feature vector gave good performance for the ringdown-only search. In both cases, we used a training set of "clean" signal designed to carefully remove contamination from glitches within the software injection-finding time window. This technique eliminated the excursion of gravitational-wave candidate coincidences from the 2σ region of the expected background at low values of inverse combined FAR as demonstrated in Fig. 2. Additionally, we examined the performance of the

RFBDT classifier in the case where the monthly analyses used background training sets from their respective months and in the case where the monthly analyses used background training sets from the entire Period 2 analysis (i.e., the full background set). We found that using the full background training set does not result in a clear sensitivity improvement unless a data quality *hveto* parameter is introduced in the feature vector.

For the IMR search, we performed a reanalysis replacing $\rho_{\mathrm{high,combined}}$ with the ranking statistic calculated by the RFBDT, M_{forest}. Comparisons with $\rho_{\mathrm{high,combined}}$ were made separately at each veto level. For EOBNRv2 waveforms, the percent improvements in VT were largest at Veto Level 2. Depending on mass bin, the VT improvements ranged from $70_{\pm 13}\%$ to $109_{\pm 11}\%$ at Veto Level 2 and from $10_{\pm 8}\%$ to $35_{\pm 7}\%$ at Veto Level 3. For IMRPhenomB waveforms, VT improvements ranged from $9_{\pm 5}\%$ to $36_{\pm 6}\%$ regardless of veto level. Additionally, we made comparisons across veto levels, using the performance of $\rho_{\mathrm{high,combined}}$ at Veto Level 2 as the standard. We found that M_{forest} at Veto Level 2 shows greater improvement and hence a more stringent upper limit than M_{forest} at Veto Level 3. This is in contrast to the better performance of $\rho_{\mathrm{high,combined}}$ at Veto Level 3 than at Veto Level 2.

For the ringdown-only search, we evaluated the performance of the RFBDT classifier using five separate ranking statistics. Comparisons were made with respect to a ringdown-only search that used the $\rho_{\mathrm{S5/S6}}$ ranking statistic [71,82,83]. The additional four searches used the M_{forest} ranking statistic for various instantiations of the RFBDT classifier. Comparisons with $\rho_{\mathrm{S5/S6}}$ were made separately at each veto level. At Veto Level 2, we found that a RFBDT classifier trained on the full background set and including the data quality *hveto* parameter in the feature vector resulted in VT improvements in the range $61_{\pm 4}\%$–$241_{\pm 12}\%$ for $q = 1$ EOBNRv2 waveforms and in the range $62_{\pm 6}\%$–$236_{\pm 14}\%$ $q = 4$ EOBNRv2 waveforms. At Veto Level 3, this same configuration resulted in VT improvements in the range $39_{\pm 4}\%$–$89_{\pm 8}\%$ for $q = 1$ EOBNRv2 waveforms and in the range $39_{\pm 5}\%$–$111_{\pm 18}\%$ $q = 4$ EOBNRv2 waveforms. Again, we made comparisons across veto levels, using the performance of $\rho_{\mathrm{S5/S6}}$ at Veto Level 2 as the standard. Unlike the IMR search, we found that M_{forest} at Veto Level 3 shows greater improvement and hence a more stringent upper limit than M_{forest} at Veto Level 2. Additionally, we found that adding an *hveto* parameter at Veto Level 2 does not result in the same increase in sensitivity obtained by applying level 3 vetoes to a search using the basic implementation of the RFBDT classifier. With more sophisticated methods for adding data quality information to the feature vector, we may see additional improvements or different behavior. Further exploration is needed.

In general, for each search, we found that the RFBDT multivariate classifier results in a considerably more sensitive search than the empirical SNR-based statistic at both veto levels. The software for constructing clean injection sets and the RFBDTs is now implemented in the LALSuite gravitational-wave data analysis routines for use with other matched-filter searches. More investigations will be needed to understand whether lower mass searches for gravitational waves from binary coalescence would benefit from the use of multivariate classification with supervised MLAs. For higher mass searches, particularly those susceptible to contamination from noise transients, RFBDT multivariate classifiers have proven to be a valuable tool for improving search sensitivity.

ACKNOWLEDGMENTS

We gratefully acknowledge the National Science Foundation for funding LIGO, and LIGO Scientific Collaboration and Virgo Collaboration for access to these data. P. T. B. and N. J. C. were supported by NSF Grant No. PHY-1306702. S. C. was supported by NSF Grants No. PHY-0970074 and No. PHY-1307429. D. T. was supported by NSF Grants No. PHY-1205952 and No. PHY-1307401. C. C. was partially supported by NSF Grants No. PHY-0903631 and No. PHY-1208881. This document has been assigned LIGO laboratory document number P1400231. The authors would like to acknowledge Thomas Dent, Chad Hanna, and Kipp Cannon for work during the initial phase of this analysis. The authors would also like to thank Alan Weinstein, Gregory Mendell, and Marco Drago for useful discussion and guidance.

APPENDIX: FEATURE VECTOR DENSITY HISTOGRAMS

Figures 8–13 show the density distributions of the feature vector statistics for both the IMR and ringdown-only searches. The distributions show the degree of separation of the simulated signal (red) and background (grey) training sets achieved by each statistic alone. The multivariate statistical classifier gives us the ability to use all these gravitational-wave data analysis statistics to calculate a combined FAR. Details on each of these statistics are given in Sec. VI C.

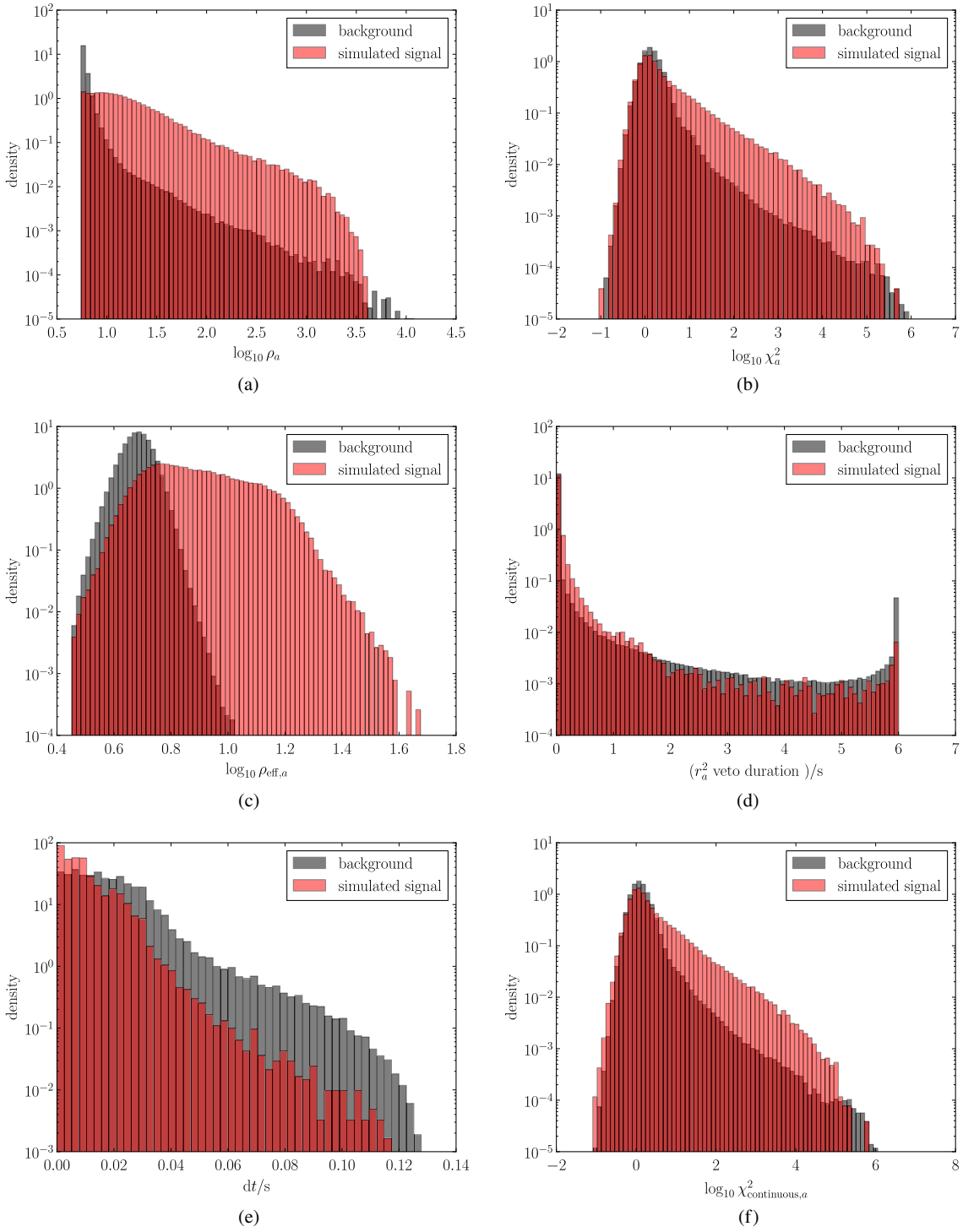


FIG. 8 (color online). Signal and background density distributions for a selection of feature vector statistics for the IMR search.

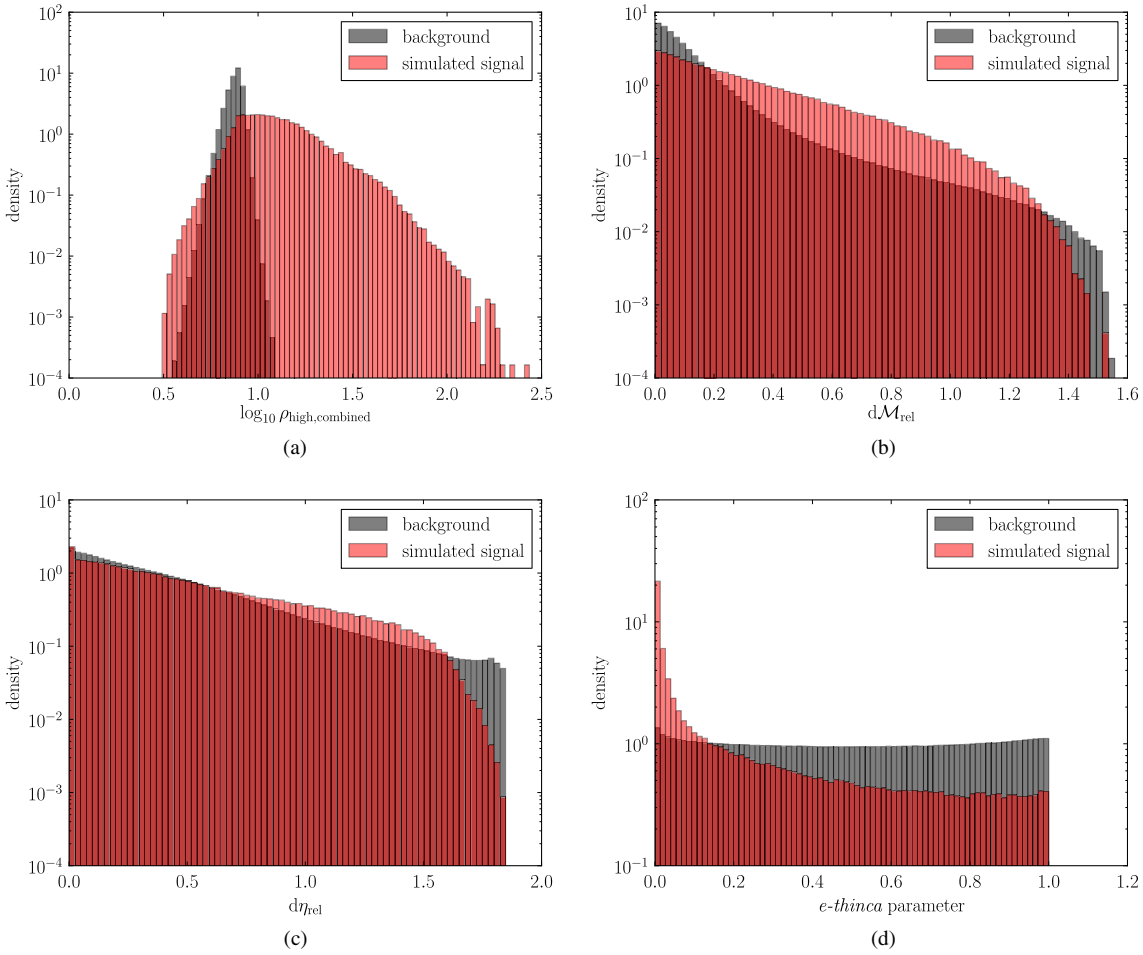


FIG. 9 (color online). Signal and background density distributions for a selection of feature vector statistics for the IMR search.

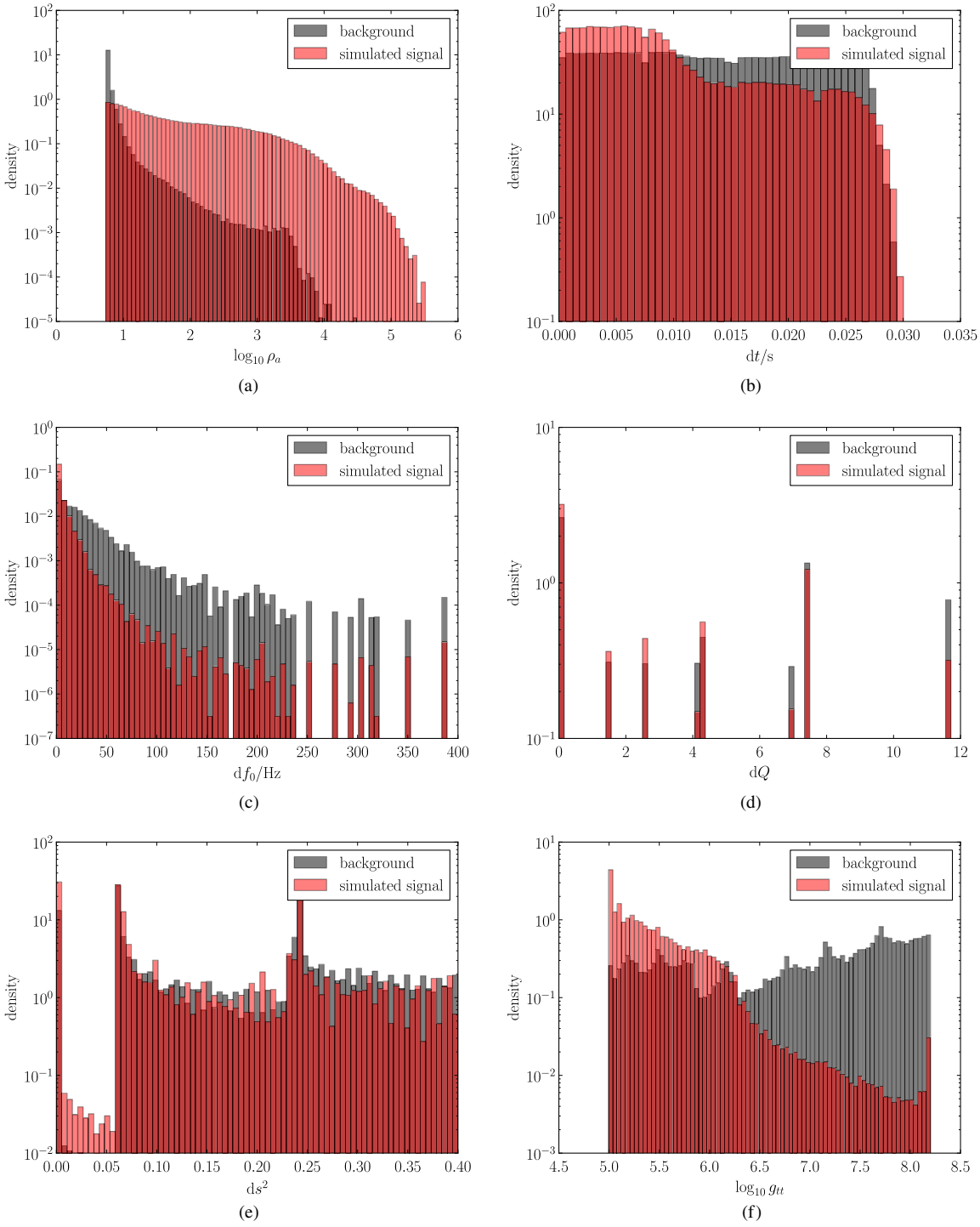


FIG. 10 (color online). Signal and background density distributions for a selection of feature vector statistics for the ringdown-only search.

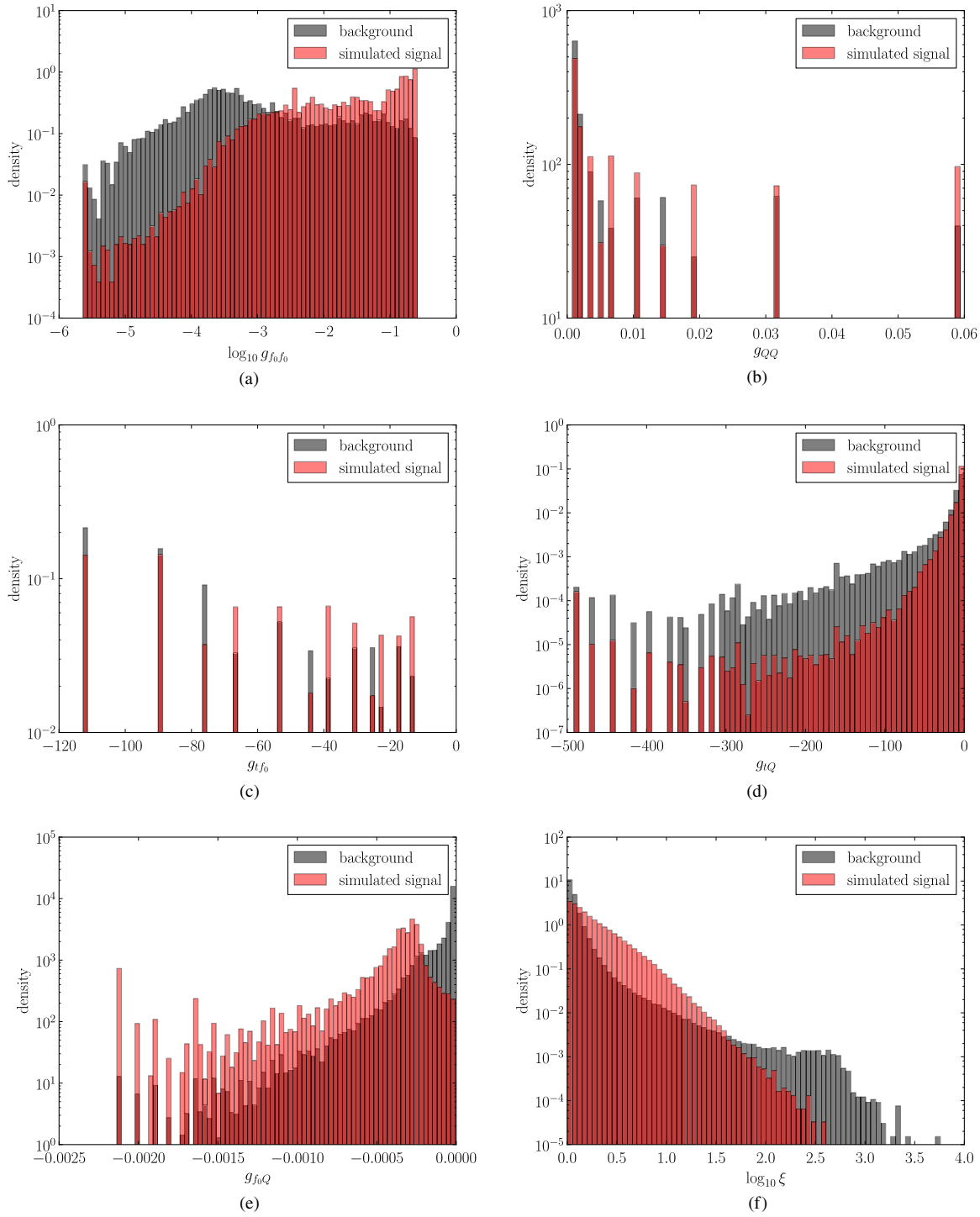


FIG. 11 (color online). Signal and background density distributions for a selection of feature vector statistics for the ringdown-only search.

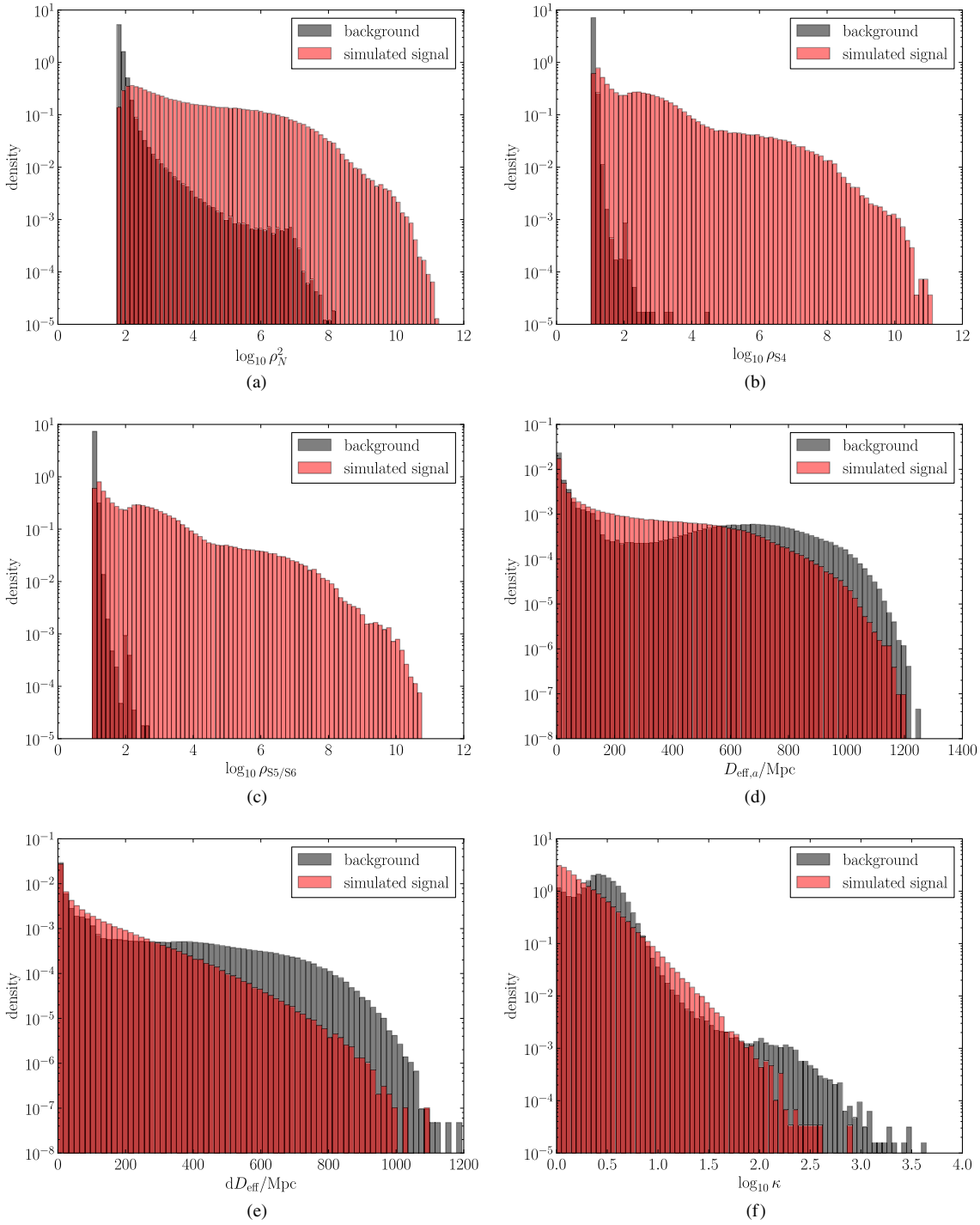


FIG. 12 (color online). Signal and background density distributions for a selection of feature vector statistics for the ringdown-only search.

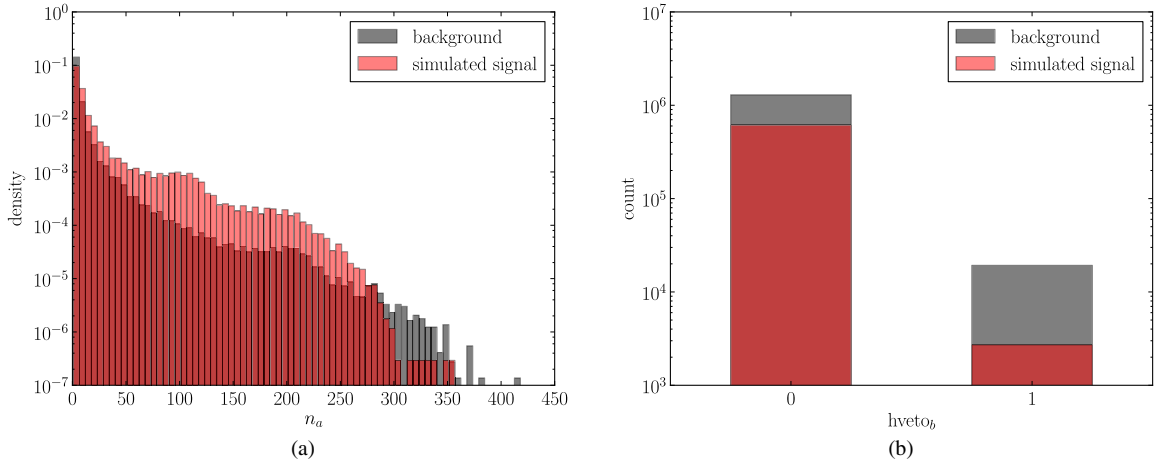


FIG. 13 (color online). Signal and background density distributions for a selection of feature vector statistics for the ringdown-only search.

[1] G. M. Harry and (LIGO Scientific Collaboration), Classical Quantum Gravity **27**, 084006 (2010).
[2] (Virgo Collaboration), https://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0027A-09.pdf (2009).
[3] K. Somiya, Classical Quantum Gravity **29**, 124007 (2012).
[4] B. Iyer, T. Souradeep, C. S. Unnikrishnan, S. Dhurandhar, S. Raja, A. Kumar, and A. Sengupta, https://dcc.ligo.org/LIGOM1100296/public (2011).
[5] J. Abadie *et al.*, Classical Quantum Gravity **27**, 173001 (2010).
[6] J. Aasi *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **87**, 022002 (2013).
[7] J. Aasi *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **89**, 102006 (2014).
[8] C. Cutler *et al.*, Phys. Rev. Lett. **70**, 2984 (1993).
[9] B. S. Sathyaprakash and B. F. Schutz, Living Rev. Relativity **12**, 2 (2009).
[10] B. P. Abbott *et al.*, Rep. Prog. Phys. **72**, 076901 (2009).
[11] B. P. Abbott *et al.*, Phys. Rev. D **79**, 122001 (2009).
[12] B. P. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **80**, 047101 (2009).
[13] J. Abadie *et al.*, Phys. Rev. D **85**, 082002 (2012).
[14] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **83**, 122005 (2011).
[15] K. C. Cannon, Classical Quantum Gravity **25**, 105024 (2008).
[16] P. T. Baker, Ph.D. thesis, Montana State University, 2013.
[17] K. A. Hodge, Ph.D. thesis, California Institute of Technology, 2014.
[18] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H. Lin, *Learning from Data: A Short Course* (AMLBook, 2012).
[19] I. Narsky and F. C. Porter, *Statistical Analysis Techniques in Particle Physics: Fit, Density Estimation, and Supervised Learning* (Wiley-VCH, New York, 2014).
[20] R. Hecht-Nielsen, in *Proceedings of International Joint Conference on Neural Networks* (IEEE, Washington, 1989), Vol. 1, pp. 593–605.
[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer Science + Business Media, New York, 2009).
[22] C. Cortes and V. Vapnik, Mach. Learn. **20**, 273 (1995).
[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, 1st ed. (Cambridge University Press, Cambridge, UK, 2000).
[24] L. Breiman, Mach. Learn. **45**, 5 (2001).
[25] W. W. Zhu *et al.*, Astrophys. J. **781**, 117 (2014).
[26] J. S. Bloom *et al.*, Publ. Astron. Soc. Pac. **124**, 1175 (2012).
[27] H. Brink, J. W. Richards, D. Poznanski, J. S. Bloom, J. Rice, S. Negahban, and M. Wainwright, Mon. Not. R. Astron. Soc. **435**, 1047 (2013).
[28] S. Whiteson and D. Whiteson, Engineering Applications of Artificial Intelligence **22**, 1203 (2009).
[29] I. Narsky, http://www.hep.caltech.edu/~narsky/SPR_Caltech_Oct2005.pdf (2005).
[30] T. S. Adams, D. Meacher, J. Clark, P. J. Sutton, G. Jones, and A. Minot, Phys. Rev. D **88**, 062006 (2013).
[31] K. Kim, I. W. Harry, K. A. Hodge, Y. Kim, C. Lee, H. K. Lee, J. J. Oh, S. H. Oh, and E. J. Son (to be published).
[32] R. Biswas *et al.*, Phys. Rev. D **88**, 062003 (2013).
[33] J. Aasi *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **112**, 131101 (2014).
[34] K. A. Postnov and L. R. Yungelson, Living Rev. Relativity **9**, 6 (2006).
[35] J. A. Faber and F. A. Rasio, Living Rev. Relativity **15**, 8 (2012).

[36] C. Helstrom, *Statistical Theory of Signal Detection*, International Series of Monographs on Electronics and Instrumentation (Pergamon Press, New York, 1960).
[37] L. A. Wainstein and V. D. Zubakov, *Extraction of Signals from Noise* (Prentice-Hall, Englewood Cliffs, NJ, 1962).
[38] R. Prix and B. Krishnan, Classical Quantum Gravity **26**, 204013 (2009).
[39] L. Blanchet, Living Rev. Relativity **5**, 3 (2002).
[40] http://www.ligo.caltech.edu/advLIGO/scripts/ref_des.shtml.
[41] B. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **72**, 082002 (2005).
[42] J. Centrella, J. G. Baker, B. J. Kelly, and J. R. van Meter, Rev. Mod. Phys. **82**, 3069 (2010).
[43] M. Hannam, Classical Quantum Gravity **26**, 114001 (2009).
[44] U. Sperhake, E. Berti, and V. Cardoso, C. R. Phys. **14**, 306 (2013).
[45] K. Kokkotas and B. Schmidt, Living Rev. Relativity **2**, 2 (1999).
[46] E. Berti, V. Cardoso, J. A. Gonzalez, U. Sperhake, M. Hannam, S. Husa, and B. Brügmann, Phys. Rev. D **76**, 064034 (2007).
[47] B. P. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **80**, 062001 (2009).
[48] L. Blanchet, T. Damour, B. R. Iyer, C. M. Will, and A. G. Wiseman, Phys. Rev. Lett. **74**, 3515 (1995).
[49] L. Blanchet, T. Damour, G. Esposito-Farèse, and B. R. Iyer, Phys. Rev. Lett. **93**, 091101 (2004).
[50] A. Buonanno, Y. Pan, J. G. Baker, J. Centrella, B. J. Kelly, S. T. McWilliams, and J. R. van Meter, Phys. Rev. D **76**, 104049 (2007).
[51] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys. Rev. D **85**, 122006 (2012).
[52] S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).
[53] A. Buonanno, G. B. Cook, and F. Pretorius, Phys. Rev. D **75**, 124018 (2007).
[54] L. Blackburn *et al.*, Classical Quantum Gravity **25**, 184004 (2008).
[55] J. Slutsky *et al.*, Classical Quantum Gravity **27**, 165023 (2010).
[56] T. Isogai (LIGO Scientific Collaboration, and Virgo Collaboration), J. Phys. Conf. Ser. **243**, 012005 (2010).
[57] N. Christensen (LIGO Scientific Collaboration, and Virgo Collaboration), Classical Quantum Gravity **27**, 194010 (2010).
[58] J. McIver, Classical Quantum Gravity **29**, 124010 (2012).
[59] J. Aasi *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Classical Quantum Gravity **29**, 155002 (2012).
[60] C. Robinson, B. Sathyaprakash, and A. S. Sengupta, Phys. Rev. D **78**, 062002 (2008).
[61] H. Nakano, H. Takahashi, H. Tagoshi, and M. Sasaki, Phys. Rev. D **68**, 102003 (2003).
[62] S. Babak *et al.*, Phys. Rev. D **87**, 024033 (2013).
[63] B. Allen, Phys. Rev. D **71**, 062001 (2005).
[64] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **84**, 124052 (2011).
[65] P. Ajith *et al.*, Phys. Rev. Lett. **106**, 241101 (2011).
[66] LIGO Scientific Collaboration and Virgo Collaboration, arXiv:1203.2674.
[67] R. Adhikari, P. Fritschel, and S. Waldman, https://dcc.ligo.org/LIGO-T060156/public (2006).
[68] T. Accadia *et al.*, Classical Quantum Gravity **28**, 025005 (2011).
[69] B. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **77**, 062002 (2008).
[70] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **83**, 122005 (2011); **86**, 069903(E) (2012).
[71] D. Talukder, S. Bose, S. Caudill, and P. T. Baker, Phys. Rev. D **88**, 122002 (2013).
[72] S. Mohapatra, L. Cadonati, S. Caudill, J. Clark, C. Hanna, S. Klimenko, C. Pankow, R. Vaulin, G. Vedovato, and S. Vitale, Phys. Rev. D **90**, 022001 (2014).
[73] D. G. Keppel, Ph.D. thesis, California Institute of Technology, 2009.
[74] I. Narsky, arXiv:physics/0507143.
[75] E. Bauer and R. Kohavi, Mach. Learn. **36**, 105 (1999).
[76] C. Gini, *Variability and Mutability*, edited by E. Pizetti and T. Salvemini (Libreria Eredi Virgilio Veschi, Rome, 1955).
[77] S. Bose, T. Dayanga, S. Ghosh, and D. Talukder, Classical Quantum Gravity **28**, 134009 (2011).
[78] A. Rodriquez, Master's thesis, Louisiana State University, 2007.
[79] C. Hanna, Ph.D. thesis, Louisiana State University, 2008.
[80] L. M. Goggin, Ph.D. thesis, California Institute of Technology, 2008.
[81] J. R. Smith, T. Abbott, E. Hirose, N. Leroy, D. MacLeod, J. McIver, P. Saulson, and P. Shawhan, Classical Quantum Gravity **28**, 235005 (2011).
[82] D. Talukder, Ph.D. thesis, Washington State University, 2012.
[83] S. E. Caudill, Ph.D. thesis, Louisiana State University, 2012.
[84] J. D. E. Creighton, Phys. Rev. D **60**, 021101 (1999).
[85] N. J. Cornish and T. B. Littenberg, arXiv:1410.3835.
[86] Z. Ivezic, A. Connolly, J. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (Princeton University Press, Princeton, NJ, 2014).
[87] E. Barausse and L. Rezzolla, Astrophys. J. **704**, L40 (2009).